Three Essays on Financial Engineering

Dissertation
submitted to the
Faculty of Business, Economics and Informatics
of the University of Zurich

to obtain the degree of
Doktor der Wirtschaftswissenschaften, Dr. oec.
(corresponds to Doctor of Philosophy, PhD)

presented by

Steven Patrick Schärer
from Möriken-Wildegg, AG

approved in July 2020 at the request of
Prof. Dr. Markus Leippold
Prof. Dr. Erich Walter Farkas

The Faculty of Business, Economics and Informatics of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, 15.07.2020

Chairman of the Doctoral Board: Prof. Dr. Steven Ongena

Acknowledgments

This thesis is the culmination of five years of research activity. It reflects how my research interests changed over time, from more theoretical to very much applied, and its completion would not have been possible without the support of a multitude of people.

First, I would like to express my deepest gratitude to Prof. Dr. Markus Leippold. He kindly accepted me as an external PhD student and was always readily available for discussions, feedback and suggestions. I have benefited tremendously from his experience in conducting research and writing research papers. His lecture on Financial Engineering was among the most rewarding during my studies and I have heavily relied on the knowledge he conveyed throughout my career so far.

I am also very grateful to Prof. Dr. Walter Farkas for generously agreeing to co-supervise this thesis. His lecture on Mathematical Foundations for Finance inspired me to switch to the Master's program in Quantitative Finance after studying Mathematics, which ultimately resulted in me pursuing this PhD.

I deeply appreciate my friends and former colleagues for keeping me grounded in the reality of how financial models are applied in practice: Francesca Birocchi, Dr. Jeremy Callner, Jacopo Conte, Krisztian Cseh, Dr. Rémi Debuquois, Dr. Tim Grob, Raluca Guran, Dr. Florian Herzog, Dr. Michael Holzhauser, Valeriya Kolesnikova, Dr. Raphaël Lamon, Marco Laube, Vivi Papoula, Dr. Sebastiano Rossi, Dr. Martin Sack, André Thea, Martin Trautmann, Damian Tschirky, Dr. Lukas Wäger and Simon Wasle.

I am ever indebted to my parents Ursula and Werner for supporting me throughout my studies. A particular thanks goes to my father for half-jokingly suggesting I study pure mathematics from age ten onwards. I ended up joining the dark side, but he meant well.

My eternal gratitude goes to my wife Jasmine and my son Livio. Without my wife's support and understanding I would have struggled tackling the most mundane of tasks, much less completing a PhD. My son, having been born in the final phase of writing this thesis, gave my life a new perspective altogether including an appreciation of working on an introduction of a PhD thesis at three o'clock in the morning. To them I dedicate this thesis.

Zurich, March 2020 Steven Schärer

Contents

Part I

Introduction

Introduction and Summary of Research Results

This thesis is a collection of three papers in the area of Financial Engineering, addressing questions in the areas of market liquidity, optimal execution and Value-at-Risk modeling.

The first research paper, *Discrete-Time Option Pricing with Stochastic Liquidity* (joint work with Markus Leippold), is concerned with the impact of market liquidity on the formation of bid and ask prices in the options market. We thereby relax one of the most fundamental assumptions made in traditional option pricing models, that buying and selling happens at the same price. This work is an extension of the Conic Finance framework developed by Cherny and D. B. Madan (2009). In that framework it is stipulated that there is a central counterparty which buys and sells from all market participants. It does so at the highest, respectively lowest price that is acceptable to it. Acceptability is defined in terms of distortion risk measures, with a single parameter describing the market liquidity, controlling the degree of distortion. Empirical studies by Corcuera et al. (2012) and Albrecher et al. (2013) suggested that a single market liquidity parameter is not enough to explain the whole surface of bid and ask prices that can be observed. To address this issue, we extend the multi-period Conic Finance model first presented in D. B. Madan (2010) to allow for a stochastic market liquidity process. In our empirical application to the S&P 500 options market we show that modeling market liquidity with a CIR process and the Conditional Value-at-Risk distortion risk measure gives remarkably good fits for bid and ask prices.

In the second research paper, *Optimal Conic Execution Strategies with Stochastic Liquidity* (joint work with Markus Leippold), we bring the Conic Finance framework to the area of optimal execution

of large positions and therefore combine two strands in the market liquidity literature that did not intersect before. When executing large buy or sell orders, market participants are generally worried about impacting prices by using up too much of the available liquidity. This concern can be alleviated by splitting up large orders and executing them over a longer period of time. However, this introduces the risk of the market moving in an adverse direction, therefore making this a trade-off between market and liquidity risk. We contribute to the existing literature by introducing a model for the bid and ask price impacted by large orders with the ideas developed in Leippold and Schärer (2017), applied to a single-step setting. We show that for the Wang distortion risk measure and constant volatility our model reduces to existing models suggested by Almgren and Chriss (2001) and Cheridito and Sepin (2014) and a numerical experiment demonstrates the versatility introduced by the Conic Finance framework.

Finally, in *Deep Learning for (Intra-Horizon) Value-at-Risk Forecasting and Application to Conditional Value-at-Risk* we study how novel deep learning methods can be used to create models for forecasting Value-at-Risk. We propose a long short-term memory (LSTM) deep neural network to forecast quantiles of return distributions over multiple time horizons and quantile levels. By training the model on a large sample of current and former S&P 500 constituents we demonstrate that it is not necessary to re-train the model to produce forecasts for time series that were not in the original training set. We also show that training the model just once on twenty years of data is enough to produce rolling forecasts over six years that are comparable or better than ones coming from state-of-the-art GARCH-based models that are re-calibrated daily. Our approach can be used to forecast intra-horizon Value-at-Risk, as introduced by Boudoukh et al. (2004), as well and also compares favorably with existing methods. We finish by motivating how a network calibrated to forecast Value-at-Risk can be used to forecast Conditional Value-at-Risk.

References

Albrecher, H., Guillaume, F., & Schoutens, W. (2013). Implied liquidity: Model sensitivity. *Journal of Empirical Finance*, *23*, 48–67.

Almgren, R., & Chriss, N. (2001). Optimal execution of portfolio transactions. *Journal of Risk*, *3*, 5–40.

Boudoukh, J., Stanton, R., Richardson, M. P., & Whitelaw, R. (2004). MaxVaR: Long-horizon value at risk in a mark-to-market environment. *Journal of Investment Management*, *2*(3), 1–6.

Cheridito, P., & Sepin, T. (2014). Optimal trade execution under stochastic volatility and liquidity. *Applied Mathematical Finance*, *21*(4), 342–362.

Cherny, A., & Madan, D. B. (2009). New measures for performance evaluation. *Review of Financial Studies*, *22*(7), 2571–2606.

Corcuera, J. M., Guillaume, F., Madan, D. B., & Schoutens, W. (2012). Implied liquidity: Towards stochastic liquidity modelling and liquidity trading. *International Journal of Portfolio Analysis and Management*, *1*(1), 80–91.

Leippold, M., & Schärer, S. (2017). Discrete-time option pricing with stochastic liquidity. *Journal of Banking & Finance*, *75*, 1–16.

Madan, D. B. (2010). Conserving capital by adjusting deltas for gamma in the presence of skewness. *Journal of Risk and Financial Management*, *3*(1), 1–25.

Part II

Research Papers

Discrete-Time Option Pricing with Stochastic Liquidity

Markus Leippold and Steven Schärer

This paper is published in:

Leippold, M., Schärer, S., 2017. Discrete-time option pricing with stochastic liquidity. Journal of Banking & Finance 75, 1-16. Copyright by Elsevier BV. DOI:
https://doi.org/10.1016/j.jbankfin.2016.11.014

Abstract

Classical option pricing theories are usually built on the law of one price, neglecting the impact of market liquidity that may contribute to significant bid-ask spreads. Within the framework of Conic Finance, we develop a stochastic liquidity model, extending the discrete-time constant liquidity model of D. B. Madan (2010). With this extension, we can replicate the term and skew structures of bid-ask spreads typically observed in option markets. We show how to implement such a stochastic liquidity model within our framework using multidimensional binomial trees and we calibrate it to call and put options on the S&P 500.

JEL Classification: C51; D52; G12; G13
Keywords: Market Liquidity; Bid-Ask Spreads; Option Pricing; Stochastic Liquidity; Conic Finance

1 Introduction

Classical option pricing theories are usually based on the paradigm of complete and frictionless markets. However, even in financial markets that are considered to be highly competitive, we do observe drops in liquidity, which in times of financial turmoils may be significant and spark concerns among market participants. Liquidity has many different facets. In this paper, we measure liquidity as the spread between bid and ask prices. Illiquid assets are characterized by a high spread. When illiquidity draws a wedge between bid and ask prices, we can no longer rely on the law of one price.

The first attempts to explain bid-ask spreads were made by introducing transaction costs such as commission charges or inventory costs.[1] However, these models often fail to explain the magnitude of the spreads observed in the markets. Especially after the financial crisis of 2008, bid-ask spreads of many assets were persistently high and at a level that cannot be explained by transaction costs alone.[2] A different approach was taken by D. B. Madan and Cherny (2010b) which is based on theory of Conic Finance, originating from the work by Cherny and D. B. Madan (2009). The basic premise is that the market takes the role of a central counterparty that buys and sells assets from and to investors. The investor buys at the ask price and sells at the bid price. The difference of these prices gives rise to the bid-ask spread observed in financial markets. The central counterparty is viewed as passive in that it does not maximize some utility function, but rather carries out all trades that are acceptable to it.[3]

D. B. Madan and Cherny (2010b) propose to model market illiquidity by a single market stress level parameter, according to which the market assigns bid and ask prices to assets based on the concept of acceptability indices. This static liquidity model was further extended and taken to the data in Corcuera et al. (2012) and Albrecher et al. (2013). These papers suggests at least two stylized facts for implied liquidity. First, market liquidity implied by real-world data exhibits both a skew and a term structure. This observation is in stark contrast to the assumption of a single liquidity parameter over all maturities and strikes. Second, they show that when we calibrate a single market

[1] See, e.g., Davis et al. (1993), Shreve and H Mete Soner (1994), Halil M Soner et al. (1995), Cvitanić and Karatzas (1996), Barles and Halil Mete Soner (1998).

[2] See, e.g., Pedersen (2009) for the average bid-ask spreads of large-cap U.S. stocks from June 2006 to June 2009.

[3] Acceptability itself is measured by acceptability indices, which are rooted in the theory of coherent risk measures as developed in Artzner, Delbaen, Eber, and Heath (1999).

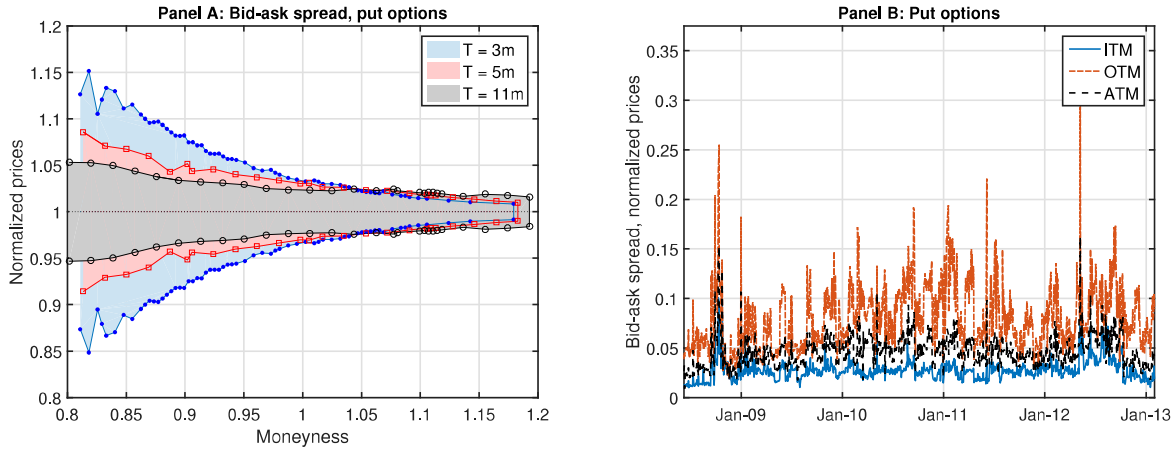


Figure 1. In Panel A, we plot the bid-ask spread for selected maturity slices of the European puts on the S&P 500 on July 20, 2012. We normalize the bid-ask spreads by the mid-prices. Moneyness is expressed as the ratio of strike over forward price. In Panel B, we plot the time-series of bid-ask spreads of in-the-money (ITM, 120% moneyness), out-of-the-money (OTM, 80% moneyness), and at-the-money (ATM, 100% moneyness) puts on the S&P500 with maturity of 5 months on July 20, 2012.

liquidity parameter for the S&P 500 option market, we obtain a time-series of the implied liquidity parameter with a mean-reverting stochastic behavior over time.

These stylized facts are illustrated in Figure 1. In Panel A, we plot the bid-ask spreads in terms of normalized prices of European puts written on the S&P 500 index. Clearly, bid-ask spreads differ across moneyness and time-to-maturity. In Panel B of Figure 1, we plot the historical bid-ask spreads for European puts with a maturity of five months. Clearly, these historical spreads change over time and exhibit some mean-reverting behavior. Hence, the empirical evidence presented in Corcuera et al. (2012) and Albrecher et al. (2013), together with the snapshot of historical bid-ask spreads in Figure 1, provides us with valuable guidance in designing a stochastic liquidity model that may account for the skew and term structure effects of implied liquidity.

We contribute to the steadily growing literature on liquidity modeling for option pricing in two ways. First, by making the setup of the discrete-time constant liquidity model of D. B. Madan (2010) more rigorous, we can simplify his results and extend the constant liquidity model to a stochastic liquidity framework. Our Theorem 1 allows us to represent bid and ask prices under stochastic liquidity given by backward recursions as time-consistent and dynamically translation invariant nonlinear expectations. This result opens the door to introduce stochastic liquidity in the

Conic Finance framework. As an illustration, we apply a specific stochastic liquidity model using multidimensional binomial trees to the S&P 500 index option market. We show that this extension improves the fit of the term and skew structures in bid-ask spreads observed in markets. To the best of our knowledge, this is the first model to treat liquidity as a separate process that can be applied to the pricing of bid-ask spreads of derivatives. Compared to other approaches, our model is also suitable for deriving the bid and ask prices of path-dependent options such as Asian and Barrier options.

There have been various other endeavors on how to introduce dynamic bid-ask spreads in option pricing based on the model of D. B. Madan and Cherny (2010b). An obvious way to do so is to model the bid and ask price as two separate stochastic processes, as suggested in D. B. Madan and Schoutens (2014). However, it can be considered a drawback that for payoffs which are not comonotone with a long or short stock position, this approach only gives lower and upper bounds for bid and ask prices. Another avenue is to follow the literature of dynamic risk measures. Mirroring the steps of the static one-step model, Bielecki, Cialenco, Iyigunler, et al. (2013) define dynamic acceptability indices with the help of dynamic coherent risk measures as discussed in, e.g., Riedel (2004) and Artzner, Delbaen, Eber, Heath, and Ku (2007). The disadvantage of using dynamic coherent risk measures is that they are not as tractable and intuitive compared to the static setup. It is furthermore not clear how a stochastic liquidity component could be incorporated. Biagini and Bion-Nadal (2014) tackle the issue in a similar way and arrive at a continuous-time version, while Bielecki, Cialenco, and Chen (2015) make use of Backward Stochastic Difference Equations (BSΔEs) and Rosazza Gianin and Sgarra (2013) derive dynamic risk measures from g-expectations.

Other than the approaches described above which are all based on or inspired by Conic Finance, there is a large body of literature that explores liquidity and bid-ask spreads in option markets.[4] One way to derive bid and ask prices of derivatives is by considering the replication costs induced by an illiquid underlying. A popular model in this direction was conceived by Çetin, R. A. Jarrow, et al. (2004) who propose to model illiquidity by assuming that prices of underlyings are provided by a stochastic supply curve, that is not impacted by the actions of buyers and sellers. The resulting

[4]See, e.g., George and Longstaff (1993), R. Engle and Neri (2010), Chou et al. (2011), Chan and Chung (2012), Bongaerts et al. (2011), P. Christoffersen, Goyenko, et al. (2018), and Feng et al. (2014), to name a few.

bid and ask prices are then dependent on the trade size which differs from our assumption of a trade-invariant bid-ask spread. They find that, in discrete time, hedging derivatives by trading the illiquid underlying incurs liquidity costs. However, results in Çetin, R. Jarrow, et al. (2006) indicate that this approach can only partially explain bid-ask spreads of derivatives observed in the market. In a separate study, Chou et al. (2011) also conclude that it is not sufficient to only consider the underlying's liquidity, but also an option's own liquidity must be taken into account. In contrast, our model does not specifically differentiate between underlying and option liquidity and indeed does not use replicating strategies to derive option prices. Hence, we assume that all information regarding liquidity is contained in the bid and ask prices of the option market.[5]

The rest of the paper is structured as follows. In Section 2, we review the one-period framework of D. B. Madan and Cherny (2010b). Section 3 introduces the multi-period model with stochastic liquidity. In Section 4, we bring our model to the data and show that the stochastic liquidity model helps to explain the skew and term structure typically observed in options' bid ask spreads. Finally, Section 5 concludes. All proofs are delegated to the appendix.

2 One-step static liquidity model

We start with a brief description of the one-step liquidity model presented in D. B. Madan and Cherny (2010b), since it builds the basis of our stochastic liquidity model presented in the subsequent section. To this end, we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and denote by $L^\infty := L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ the space of all essentially bounded and $\mathbb{R}$-valued random variables on $(\Omega, \mathcal{F}, \mathbb{P})$.[6] By $\mathbb{P}$, we denote the reference probability measure that we assume to be a risk-neutral measure. In a complete market, this probability measure is unique and the price of an asset is given by the $\mathbb{P}$-expectation of its future discounted cash flows, say $Q \in L^\infty$. When market incompleteness drives a wedge between bid and ask prices, we can interpret the bid (ask) price as being caused by an overweighting (underweighting) of losses and an underweighting (overweighting) of gains relative to the measure $\mathbb{P}$. Hence, we can

[5] In our approach, we cannot differentiate between how much of the illiquidity reflected in the options' bid-ask spreads are due to the illiquidity of the underlying market and how much is due to the option market itself.

[6] This choice is for simplicity only and the results can be generalized to L^p spaces with $1 \leq p < +\infty$. The discrete-time extension will only consider finite spaces, where L^p spaces are equivalent anyway.

model this weighting scheme as a distortion to the reference probability measure. For this purpose, we define a distortion function as follows.

Definition 1 (Distortion function)**.** *A function $\psi : [0,1] \longrightarrow [0,1]$ is a distortion function if and only if it is monotone, $\psi(0) = 0$ and $\psi(1) = 1$.*

The distorted probability measure $\psi \circ \mathbb{P}$ is no longer a probability measure in general. It is, however, still a finite monotone set function that is submodular, if the distortion function is concave. It is therefore possible to define a risk measure based on distorted probabilities using Choquet integrals.[7]

Definition 2 (Distortion risk measure)**.** *Let ψ be a concave distortion function and Q a future discounted cash flow. The function $\varrho^{\psi} : L^{\infty} \longrightarrow \mathbb{R}$ given by*

$$\varrho^{\psi}(Q) := \int_{-\infty}^{0} \psi(\mathbb{P}(Q \leq x))dx - \int_{0}^{\infty} (1 - \psi(\mathbb{P}(Q \leq x)))dx, \quad \forall Q \in L^{\infty}, \tag{2.1}$$

is called a distortion risk measure induced by ψ.

From the properties of the Choquet integral and because $\psi \circ \mathbb{P}$ is submodular, the function ϱ^{ψ} defined by equation (2.1) is monotone, positively homogeneous, translation invariant, and subadditive. Hence, it is a coherent risk measure. By inverting the sign of ϱ^{ψ} we obtain what is called a distorted expectation, corresponding to the intuition of weighting losses and gains differently compared to $\mathbb{P}$. In particular, we call the function $\mathbb{E}^{\psi}[\cdot] : L^{\infty} \longrightarrow \mathbb{R}$ given by

$$\mathbb{E}^{\psi}[Q] := -\varrho^{\psi}(Q), \quad \forall Q \in L^{\infty}, \tag{2.2}$$

the distorted expectation induced by ψ.

Just as coherent risk measures, $\mathbb{E}^{\psi}[\cdot]$ is nonlinear, i.e., in general $\mathbb{E}^{\psi}[Q^1+Q^2] \neq \mathbb{E}^{\psi}[Q^1]+\mathbb{E}^{\psi}[Q^2]$ for $Q^1, Q^2 \in L^{\infty}$. Nevertheless, $\mathbb{E}^{\psi}[\cdot]$ shares many other properties with the usual expectation operator such as monotonicity, positive homogeneity, and translation invariance. The concept of nonlinear expectations, albeit in a multi-period setting, will be the cornerstone of our stochastic liquidity model.

[7] See Choquet (1953). For more details on the properties of Choquet integrals, see the standard book of Denneberg (1994).

In particular, from Definition 2 it follows that for $Q \in L^\infty$ and ψ a concave distortion function,

$$\mathbb{E}^\psi[Q] \leq \mathbb{E}[Q] \leq -\mathbb{E}^\psi[-Q], \tag{2.3}$$

where $\mathbb{E}[\cdot]$ denotes the expectation with respect to the reference pricing measure $\mathbb{P}$. Hence, distorted expectations provide an intuitive basis for the modeling of bid-ask spreads.

To measure the degree of distortion applied to the reference probability measure $\mathbb{P}$, the central counterparty is assumed to have not only one concave distortion function with which it evaluates potential trades, but a whole family $\Psi = (\psi^z)_{z\geq 0}$. This family of concave distortion functions is pointwise increasing in z in that $\psi^{z_1}(\cdot) \leq \psi^{z_2}(\cdot)$ if and only if $z_1 \leq z_2$. Additionally, ψ^0 is assumed to be the identity function. For a given distortion function $\psi^\gamma \in \Psi$ under which cash flows are evaluated, we can interpret γ as the market liquidity level. The more illiquid the market becomes (i.e., the higher γ is), the more distorted the reference probability measure becomes. A liquidity level of zero implies that no distortion at all is applied, which corresponds to perfect liquidity and hence to the complete market case, in which the law of one price holds.[8]

A discounted cash flow $Q \in L^\infty$ is deemed acceptable at γ if and only if $\mathbb{E}^{\psi^\gamma}[Q] \geq 0$. The market is assumed to competitively execute only acceptable trades. Therefore, the bid price $b^{\Psi,\gamma}(Q)$ of a discounted cash flow $Q \in L^\infty$ is the highest price the market is willing to pay for the net position to be acceptable according to the market liquidity level, i.e.,

$$b^{\Psi,\gamma}(Q) := \sup\{b \in \mathbb{R} \,|\, \mathbb{E}^{\psi^\gamma}[Q - b] \geq 0\} = \mathbb{E}^{\psi^\gamma}[Q], \tag{2.4}$$

where the last equality follows from the translation invariance of coherent risk measures. A similar argument can be made for the ask price and leads to the following definition of bid and ask prices.[9]

Definition 3 (Single-period bid-ask prices)**.** *Let $\Psi = (\psi^z)_{z\geq 0}$ be a pointwise increasing family of concave distortion functions and $\gamma > 0$ the market liquidity level. Then, the ask price of a discounted*

[8]The liquidity measure in our model is not directly defined by observable variables in the market such as trading volume or the bid-ask spreads. Instead, it is inferred from a comparison of market and model implied bid-ask spreads.

[9]This approach differs significantly from deriving bid and ask prices of derivatives based on a replicating trading strategy as in, e.g., Çetin, R. A. Jarrow, et al. (2004).

cash flow $Q \in L^\infty$ is given by

$$a^{\Psi,\gamma}(Q) := -\mathbb{E}^{\psi^\gamma}[-Q] = \int_{-\infty}^{0} (\psi^\gamma(\mathbb{P}(Q > x)) - 1)\, dx + \int_{0}^{\infty} \psi^\gamma(\mathbb{P}(Q > x))dx, \tag{2.5}$$

and its bid price is

$$b^{\Psi,\gamma}(Q) := \mathbb{E}^{\psi^\gamma}[Q] = -\int_{-\infty}^{0} \psi^\gamma(\mathbb{P}(Q \leq x))dx + \int_{0}^{\infty} (1 - \psi^\gamma(\mathbb{P}(Q \leq x)))\, dx. \tag{2.6}$$

Without any further assumptions, the bid price is always less than or equal to the ask price. Bid and ask price also envelop the undistorted price, i.e.,

$$b^{\Psi,\gamma}(Q) \leq \mathbb{E}[Q] \leq a^{\Psi,\gamma}(Q) \quad \forall Q \in L^\infty. \tag{2.7}$$

D. B. Madan and Cherny (2010b) derive these bid and ask prices from the theory of acceptability indices, which are functions $\alpha : L^\infty \to [0, \infty]$. In particular, they call a net cash flow, or trade, $\widetilde{Q} \in L^\infty$ acceptable at a certain market liquidity level γ if and only if $\alpha(\widetilde{Q}) \geq \gamma$.[10] Cherny and D. B. Madan (2009) show that acceptability indices can be represented by a family of coherent risk measures $(\varrho^z)_{z\geq 0}$ that are continuous from above and pointwise increasing in z. For a pointwise increasing set of concave distortion functions $\Psi = (\psi^z)_{z\geq 0}$, such a family is given by the corresponding distortion risk measures $\Phi = (\varrho^{\psi^z})_{z\geq 0}$. To see that Φ is pointwise increasing in z, note that distortion risk measures retain the ordering of the associated distortion function. Furthermore, all distortion risk measures are continuous from above.[11] Hence there exists an acceptability index which corresponds to Φ.

For our model, we assume the Conditional Value-at-Risk (CVaR) as distortion measure.[12] According to, e.g., Föllmer and Schied (2011), the CVaR at level $\alpha \in (0, 1]$ can be expressed as a distortion

[10] Equivalently, $a^{\Psi,\gamma}(Q) = \sup_{\tilde{P}\in\mathcal{D}} E^{\tilde{P}}[Q]$, where $\mathcal{D}$ is the convex set of risk measures that are equivalent to $\mathbb{P}$ and are determined by α and γ.

[11] See, e.g., Föllmer and Schied (2011).

[12] Other measures that have been applied in the literature are the Wang distortion function (S. S. Wang (2000)), the MinMaxVar-distortion function introduced in Cherny and D. B. Madan (2009), or the EssSupExp-distortion function proposed by Bannör and Scherer (2014).

risk measure

$$\mathrm{CVaR}_\alpha(Q) = \varrho^{\widetilde{\psi}^\alpha}(Q) = \int_{-\infty}^{0} \widetilde{\psi}^\alpha(\mathbb{P}(Q \leq x))dx - \int_0^\infty (1 - \widetilde{\psi}^\alpha(\mathbb{P}(Q \leq x)))dx, \quad \forall Q \in L^\infty, \tag{2.8}$$

induced by the concave distortion function

$$\widetilde{\psi}^\alpha(u) := \min\left\{\frac{u}{\alpha}, 1\right\}, \quad \forall u \in [0,1]. \tag{2.9}$$

The distortion functions in our setup have to depend on a parameter taking values in $[0, \infty)$. However, the quantile parameter α is defined on $(0, 1]$. We therefore first use the change of variables $x \mapsto 1 - x$ to map $(0, 1]$ to $[0, 1)$ and then apply a sigmoid function, e.g.,

$$\varphi(x) := \frac{x}{\sqrt{1 + x^2}}, \quad \forall x \in \mathbb{R}, \tag{2.10}$$

which bijectively maps values from $[0, \infty)$ to $[0, 1)$. With that we get, for $z \geq 0$ and $\alpha = 1 - \varphi(z) \in (0, 1]$,

$$\mathrm{CVaR}_\alpha(Q) = \varrho^{\psi_{\mathrm{CVaR}}^{\varphi^{-1}(1-\alpha)}}(Q) = \varrho^{\psi_{\mathrm{CVaR}}^z}(Q) = \mathrm{CVaR}_{1-\varphi(z)}(Q), \quad \forall Q \in L^\infty, \tag{2.11}$$

for the modified concave distortion function

$$\psi_{\mathrm{CVaR}}^z(u) := \min\left\{\frac{u}{1 - \varphi(z)}, 1\right\}, \quad \forall u \in [0,1]. \tag{2.12}$$

As required, the family $\Gamma^{\mathrm{CVaR}} = (\psi_{\mathrm{CVaR}}^z)_{z \geq 0}$ is pointwise increasing in z and ψ_{CVaR}^0 is the identity function.

3 Discrete-time stochastic liquidity model

The model presented in the last section only considers a single time step. Hence, it is of limited practical value. In this section, we extend the model of D. B. Madan (2010) to treat market liquidity as a stochastic process instead of a constant to account for the stylized facts of bid and ask spreads. As in the static model, it is our goal to have bid and ask prices that are represented by nonlinear expectations. Due to the multi-period setup, we additionally require that they behave consistently

over time. These nonlinear expectations can also be used to define dynamic risk measures as we show in Remark 1.

The proof of the main theorem, which we present in Appendix A, relies on the theory of BSΔEs of Cohen and Elliott (2010b), the discrete-time analogue of Backward Stochastic Differential Equations (BSDEs) as developed by El Karoui et al. (1997). Similar to how Coquet et al. (2002) linked g-expectations to solutions of BSDEs, one can show that certain types of nonlinear expectations are solutions to BSΔEs. As in the static setup, these will be used to define the bid and ask prices of discounted cash flows.

Before we present the main result, we introduce some additional notation. We denote by $T > 0$ maturity and assume we have time points $0 = t_0 < \cdots < t_K = T$ for $K > 0$. By $\mathcal{T}_i^j := \bigcup_{i \leq l \leq j} \{t_l\}$ we denote the set of time points from t_i to t_j, where $0 \leq i, j \leq K$. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in \mathcal{T}_0^K}, \mathbb{P})$ be a finite filtered probability space satisfying the usual conditions. The underlying price $S = (S_t)_{t \in \mathcal{T}_0^K}$ is modeled as a positive but otherwise general finite state price process on the probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in \mathcal{T}_0^K}, \mathbb{P})$. As in the static setting, $\mathbb{P}$ denotes the reference probability measure that is assumed to be risk-neutral.

3.1 Time-consistent nonlinear expectations

In the spirit of Peng (2007), dynamic nonlinear expectations are defined as follows.

Definition 4 (Time-consistent nonlinear expectation)**.** *A family of functions $\mathcal{E}(\cdot \,|\, \mathcal{F}_t) : L^1(\mathcal{F}_T) \longrightarrow L^1(\mathcal{F}_t)$ for $t \in \mathcal{T}_0^K$ is a time-consistent nonlinear expectation if it satisfies, for all $s \in \mathcal{T}_0^K$ and $Q, Q^1, Q^2 \in L^1(\mathcal{F}_T)$,*

(i) monotonicity, i.e., if $Q^1 \leq Q^2$ $\mathbb{P}$-a.s., then

$$\mathcal{E}(Q^1 \,|\, \mathcal{F}_t) \leq \mathcal{E}(Q^2 \,|\, \mathcal{F}_t) \quad \mathbb{P}\text{-a.s} \tag{3.1}$$

Additionally, for $Q^1 \leq Q^2$ $\mathbb{P}$-a.s., equality holds if and only if $\mathbb{P}$-a.s. $Q^1 = Q^2$.

(ii) adaptability, i.e., $\mathcal{E}(Q \,|\, \mathcal{F}_t) = Q$ if Q is $\mathcal{F}_t$-measurable.

(iii) dynamic consistency, i.e., $\mathcal{E}(\mathcal{E}(Q \,|\, \mathcal{F}_t) \,|\, \mathcal{F}_s) = \mathcal{E}(Q \,|\, \mathcal{F}_s)$, $\mathbb{P}$-a.s. $\forall s \leq t$.

(iv) relevance, i.e., $\mathbb{1}_A \mathcal{E}(Q \,|\, \mathcal{F}_t) = \mathcal{E}(\mathbb{1}_A Q \,|\, \mathcal{F}_t)$, $\mathbb{P}$*-a.s.* $\forall A \in \mathcal{F}_t$.

The monotonicity property states that of two different payoffs, the one that is $\mathbb{P}$-a.s. smaller also has a smaller expectation under $\mathcal{E}(\cdot \,|\, \mathcal{F}_t)$ for all $t \in \mathcal{T}_0^K$. Adaptability is assumed due to our multi-period setup. Dynamic consistency is what is known as the tower property for usual conditional expectations. For nonlinear expectations, we have to explicitly make this assumption to ensure that different time steps are linked consistently. Finally, "relevance" means that at time t, the investor knows whether the underlying's path is in $A \in \mathcal{F}_t$. If this is the case, the nonlinear expectation of $\mathbb{1}_A Q$ is the same as the one of Q. Otherwise, it is zero.

Definition 5 (Dynamic translation invariance)**.** *A time-consistent nonlinear expectation, denoted by* $(\mathcal{E}(\cdot \,|\, \mathcal{F}_t))_{t \in \mathcal{T}_0^K}$*, is dynamically translation invariant if and only if for all* $t \in \mathcal{T}_0^K$*,*

$$\mathcal{E}(Q + q \,|\, \mathcal{F}_t) = \mathcal{E}(Q \,|\, \mathcal{F}_t) + q, \quad \forall Q \in L^1(\mathcal{F}_T), q \in L^1(\mathcal{F}_t). \tag{3.2}$$

Dynamic translation invariance is not generally required of nonlinear expectations. Nevertheless, we want to ensure that, e.g., a portfolio consisting of one derivative and some cash has the same bid price as adding the cash to the bid price of the derivative on its own.

3.2 Distorted conditional expectation

Instead of considering only a constant market liquidity level γ as in D. B. Madan (2010), we introduce a stochastic process to model market liquidity through time and states.

Definition 6 (Market liquidity process)**.** *The time- and state-dependent process* $\Gamma = (\gamma_t)_{t \in \mathcal{T}_0^K}$*, for* $\gamma_t \in L^1(\mathcal{F}_t)$ *and* $\gamma_t \geq 0$ $\mathbb{P}$*-a.s. for all* $t \in \mathcal{T}_0^K$ *is called market liquidity process.*

Before, we used the liquidity level $\gamma \geq 0$ to determine the degree of distortion applied to $\mathbb{P}$ by choosing ψ^γ from a family of pointwise increasing concave distortion functions Ψ. We will continue in this spirit, but since the liquidity level at any time is now a random variable, we introduce a state-dependent distortion function.

Definition 7 (Concave state-dependent distortion function)**.** *A function* $\psi : \Omega \times [0,1] \longrightarrow [0,1]$ *is*

called a concave state-dependent distortion function if and only if for all $\omega \in \Omega$, $\psi(\omega, \cdot)$ *is a concave distortion function.*

Slightly abusing notation, we will denote by $\Psi := (\psi^z)_{z \geq 0}$ the usual family of concave distortion functions that are pointwise increasing in z and define, for a random variable γ,

$$\psi^\gamma(\omega, u) := \psi^{\gamma(\omega)}(u) \quad \forall \omega \in \Omega, u \in [0,1]. \tag{3.3}$$

This gives a concave state-dependent distortion function as defined above. As such, we will continue working with the same family of concave distortion functions as in the previous section. The notion of distorted expectations is also extended to state-dependent distortion functions. Furthermore, they are now conditional on the filtration.

Definition 8 (Distorted conditional expectation)**.** *Let* ψ *be a concave state-dependent distortion function. The family of functions with elements* $\mathbb{E}_t^\psi[\cdot] : L^1(\mathcal{F}_T) \longrightarrow L^1(\mathcal{F}_t)$ *for* $t \in \mathcal{T}_0^K$*, defined* $\forall Q \in L^1(\mathcal{F}_T)$ *and* $\forall \omega \in \Omega$ *as*

$$\mathbb{E}_t^\psi[Q](\omega) := -\int_{-\infty}^0 \psi(\omega, \mathbb{P}_t(Q \leq x)(\omega))dx + \int_0^\infty 1 - \psi(\omega, \mathbb{P}_t(Q \leq x)(\omega))dx \tag{3.4}$$

is called distorted conditional expectation.

For brevity, we denote the conditional probability by[13]

$$\mathbb{P}_t(A) := \mathbb{P}(A \,|\, \mathcal{F}_t) = \mathbb{E}[\mathbb{1}_A \,|\, \mathcal{F}_t], \quad \forall A \in \mathcal{F}_T. \tag{3.5}$$

Since $\mathbb{P}_t(\cdot)(\omega)$ is a probability measure for any state $\omega \in \Omega$ and $t \in \mathcal{T}_0^K$, $\mathbb{E}_t^\psi[\cdot](\omega)$ is a distorted expectation as in the static framework. As in the static case it holds that

$$\mathbb{E}_t^\psi[Q] \leq \mathbb{E}[Q \,|\, \mathcal{F}_t] \leq -\mathbb{E}_t^\psi[-Q], \quad \mathbb{P}\text{-a.s.}, \tag{3.6}$$

for all $Q \in L^1(\mathcal{F}_T)$, $t \in \mathcal{T}_0^K$, and all concave state-dependent distortion functions ψ.

[13]Note that we assume enough regularity on $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in \mathcal{T}_0^K}, \mathbb{P})$ that a regular conditional distribution can be constructed.

3.3 Time-consistent bid and ask prices

To define the bid and ask prices in our multiperiod setting, we borrow from the one-step static liquidity model of D. B. Madan and Cherny (2010b). We start at the final maturity T and recursively apply the conditional distorted expectation. The stochastic liquidity component thereby determines the re-weighting of the reference probability measure depending on the current state and time step. Working backwards, we arrive at time t_0 and obtain today's bid and ask prices.

Definition 9 (Multi-period bid-ask prices)**.** *For $\Psi := (\psi^z)_{z\geq 0}$ a family of concave distortion functions that are pointwise increasing in z, market liquidity process $\Gamma = (\gamma_t)_{t\in\mathcal{T}_0^K}$ and $t_k \in \mathcal{T}_0^K$, the bid price of a future discounted cash flow $Q \in L^1(\mathcal{F}_T)$ at time t_k is defined as*

$$b_{t_k}^{\Psi,\Gamma}(Q) := \mathbb{E}_{t_k}^{\psi^{\gamma_{t_k}}}[\mathbb{E}_{t_{k+1}}^{\psi^{\gamma_{t_{k+1}}}}[\cdots \mathbb{E}_{t_{K-1}}^{\psi^{\gamma_{t_{K-1}}}}[Q]]]. \tag{3.7}$$

Its ask price at time t_k is given by

$$a_{t_k}^{\Psi,\Gamma}(Q) := -\mathbb{E}_{t_k}^{\psi^{\gamma_{t_k}}}[\mathbb{E}_{t_{k+1}}^{\psi^{\gamma_{t_{k+1}}}}[\cdots \mathbb{E}_{t_{K-1}}^{\psi^{\gamma_{t_{K-1}}}}[-Q]]]. \tag{3.8}$$

The distorted conditional expectation does not satisfy the tower property. Hence, it is not a time-consistent nonlinear expectation. Therefore, $b_t^{\Psi,\Gamma}(Q) \neq \mathbb{E}_t^{\psi^\gamma}[Q]$ in general. However, it is still adapted, monotone and dynamically translation invariant, as also Lemma 2 shows. By accounting for the dynamic translation invariance of the driver in the BSΔE, we are able to simplify the result in D. B. Madan (2010) and prove that the bid and ask prices from Definition 9 are time-consistent and dynamically translation invariant nonlinear expectations.[14] At the same time, we correct an error in the proof of D. B. Madan (2010), see Remark 2.

Theorem 1. *Let $\Psi := (\psi^z)_{z\geq 0}$ a family of concave distortion functions that are pointwise increasing in z and $\Gamma = (\gamma_t)_{t\in\mathcal{T}_0^K}$ a market liquidity process. Then, the bid and ask prices $(b_t^{\Psi,\Gamma})_{t\in\mathcal{T}_0^K}$ and $(a_t^{\Psi,\Gamma})_{t\in\mathcal{T}_0^K}$ are time-consistent and dynamically translation invariant nonlinear expectations.*

[14] Time-consistency of bid and ask prices prevents round-trip arbitrage opportunities. In particular, buying Q at time t_0 has to cost the same as buying a newly introduced asset that pays the ask price of Q at time t_1. This payoff could then be used to buy the asset at time t_1, after which the two approaches are equivalent.

Constant liquidity model

As a first application of Theorem 1, we can consider a constant liquidity model. In particular, by fixing $\gamma > 0$ to a constant, we obtain the model suggested by D. B. Madan (2010) albeit with simplified expressions for the bid and ask price. Assuming a recombining binomial tree model for the stock price process $(S_{t_k})_{k=0}^K$ with a constant time step size $h := \frac{T}{K}$, and $k = 1, ..., K$,

$$S_{t_k} := S_0 u^{\sum_{i=1}^k \xi_i} d^{k - \sum_{i=1}^k \xi_i}, \tag{3.9}$$

where $S_0 > 0$, $u \geq 1$, $d = 1/u$ and $(\xi_i)_{i=1}^K$ are iid random variables with values in $\{0, 1\}$, for which we set the up and down probabilities $p_u := \mathbb{P}(\xi_i = 1)$, $p_d := \mathbb{P}(\xi_i = 0)$. We also have a family of concave distortion functions Ψ that are pointwise increasing. For option payoffs that are path-dependent, it is necessary to go through the tree backward recursively to calculate the bid and ask prices at time zero. For other payoffs such as European vanilla contracts, it is possible to derive closed-form expressions. Given Theorem 1, we obtain the following analytical formulas for the bid and ask prices of European claims.

Corollary 1. *Let the future discounted cash flow be given by a function H such that $Q = H(T, S_T)$. Further assume that H is non-negative and monotonically increasing with S_T (e.g., a European call option). Then,*

$$a_0^{\Psi,\gamma}(Q) = \sum_{i=0}^K \binom{K}{i} \psi^\gamma(p_u)^i (1 - \psi^\gamma(p_u))^{K-i} H(T, S_0 u^i d^{K-i}) \tag{3.10}$$

and

$$b_0^{\Psi,\gamma}(Q) = \sum_{i=0}^K \binom{K}{i} (1 - \psi^\gamma(p_d))^i \psi^\gamma(p_d)^{K-i} H(T, S_0 u^i d^{K-i}). \tag{3.11}$$

Similarly, for a non-negative but monotonically decreasing payoff (e.g., a European put option), we have

$$a_0^{\Psi,\gamma}(Q) = \sum_{i=0}^K \binom{K}{i} \psi^\gamma(p_d)^i (1 - \psi^\gamma(p_d))^{K-i} H(T, S_0 u^{K-i} d^i) \tag{3.12}$$

and

$$b_0^{\Psi,\gamma}(Q) = \sum_{i=0}^K \binom{K}{i} (1 - \psi^\gamma(p_u))^i \psi^\gamma(p_u)^{K-i} H(T, S_0 u^{K-i} d^i). \tag{3.13}$$

The derivation follows the same steps as described in the next section.

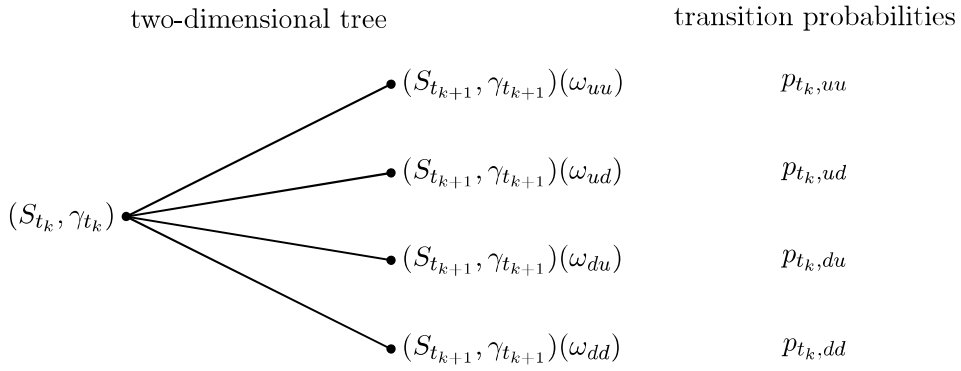


Figure 2. A single node of the two-dimensional binomial tree for (S, γ) with the corresponding transition probabilities.

Stochastic liquidity model

To generalize the previous model to a stochastic liquidity model, we proceed as follows. We describe the dynamics of S and γ by a two-dimensional recombining binomial tree $(S_t, \gamma_t, p_t, q_t)_{t \in \mathcal{T}_0^K}$ as illustrated in Figure 2. For any node (S_{t_k}, γ_{t_k}), there are four connected nodes $(S_{t_{k+1}}, \gamma_{t_{k+1}})(\omega_{uu})$, $(S_{t_{k+1}}, \gamma_{t_{k+1}})(\omega_{ud})$, $(S_{t_{k+1}}, \gamma_{t_{k+1}})(\omega_{du})$ and $(S_{t_{k+1}}, \gamma_{t_{k+1}})(\omega_{dd})$ with corresponding transition probabilities $p_{t_k,uu}$, $p_{t_k,ud}$, $p_{t_k,du}$ and $p_{t_k,dd}$. State ω_{ud} stands for an up-move of the underlying S and a down-move of the liquidity process γ. The other states and transition probabilities are denoted using the same notation.

We are now interested in calculating the bid and ask price of a future discounted cash flow $Q \in L^1(\mathcal{F}_T)$ occurring at time T. For simplicity, we restrict ourselves to non-negative Q. Since γ is non-constant, we cannot derive closed-form formulas for the bid and ask price as in the constant liquidity case. We therefore have to calculate them backward recursively according to Definition 9. To that end, we define $b^{\Psi,\Gamma}_{t_K}(Q) = a^{\Psi,\Gamma}_{t_K}(Q) = Q$ with $t_K = T$. Assuming we have already calculated $b^{\Psi,\Gamma}_{t_{k+1}}(Q)$ and $a^{\Psi,\Gamma}_{t_{k+1}}(Q)$, we can calculate bid and ask prices for every node of the previous time-step as follows. First, starting from a node (S_{t_k}, γ_{t_k}), we sort the four possible states ω_{uu}, ω_{ud}, ω_{du}, and ω_{dd} such that

$$b^{\Psi,\Gamma}_{t_{k+1}}(Q)(\omega^b_1) \geq b^{\Psi,\Gamma}_{t_{k+1}}(Q)(\omega^b_2) \geq b^{\Psi,\Gamma}_{t_{k+1}}(Q)(\omega^b_3) \geq b^{\Psi,\Gamma}_{t_{k+1}}(Q)(\omega^b_4) \tag{3.14}$$

and

$$a^{\Psi,\Gamma}_{t_{k+1}}(Q)(\omega^a_1) \geq a^{\Psi,\Gamma}_{t_{k+1}}(Q)(\omega^a_2) \geq a^{\Psi,\Gamma}_{t_{k+1}}(Q)(\omega^a_3) \geq a^{\Psi,\Gamma}_{t_{k+1}}(Q)(\omega^a_4), \tag{3.15}$$

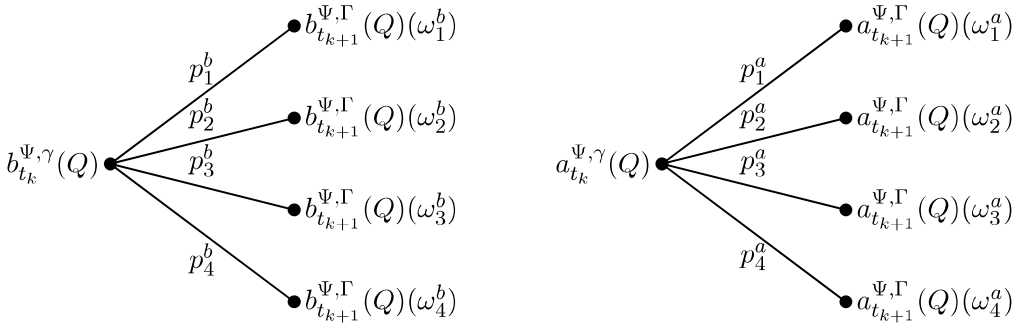


Figure 3. The same node but now the states and transition probabilities are sorted such that the bid and ask prices are highest for ω^b_1 respectively ω^a_1 and lowest for ω^b_4 respectively ω^a_4.

for states $\omega^b_i, \omega^a_i \in \{\omega_{uu}, \omega_{ud}, \omega_{du}, \omega_{dd}\}$, $i = 1, ..., 4$. In Figure 3, we plot the node for the bid and ask price. To simplify notation, we denote the transition probabilities corresponding to the states with the same sub- and superscripts. Then, since Q is non-negative, the bid price is given by

$$\begin{aligned} b^{\Psi,\Gamma}_{t_k}(Q) = \mathbb{E}^{\psi^{\gamma_{t_k}}}_{t_k}[b^{\Psi,\Gamma}_{t_{k+1}}(Q)] &= b^{\Psi,\Gamma}_{t_{k+1}}(Q)(\omega^b_1)(1-\psi^{\gamma_{t_k}}(p^b_2+p^b_3+p^b_4)) \\ &+ b^{\Psi,\Gamma}_{t_{k+1}}(Q)(\omega^b_2)(\psi^{\gamma_{t_k}}(p^b_2+p^b_3+p^b_4)-\psi^{\gamma_{t_k}}(p^b_3+p^b_4)) \\ &+ b^{\Psi,\Gamma}_{t_{k+1}}(Q)(\omega^b_3)(\psi^{\gamma_{t_k}}(p^b_3+p^b_4)-\psi^{\gamma_{t_k}}(p^b_4)) \\ &+ b^{\Psi,\Gamma}_{t_{k+1}}(Q)(\omega^b_4)\psi^{\gamma_{t_k}}(p^b_4) \end{aligned} \tag{3.16}$$

and the ask price is

$$\begin{aligned} a^{\Psi,\Gamma}_{t_k}(Q) = -\mathbb{E}^{\psi^{\gamma_{t_k}}}_{t_k}[-a^{\Psi,\Gamma}_{t_{k+1}}(Q)] &= a^{\Psi,\Gamma}_{t_{k+1}}(Q)(\omega^a_1)\psi^{\gamma_{t_k}}(p^a_1) \\ &+ a^{\Psi,\Gamma}_{t_{k+1}}(Q)(\omega^a_2)(\psi^{\gamma_{t_k}}(p^a_1+p^a_2)-\psi^{\gamma_{t_k}}(p^a_1)) \\ &+ a^{\Psi,\Gamma}_{t_{k+1}}(Q)(\omega^a_3)(\psi^{\gamma_{t_k}}(p^a_1+p^a_2+p^a_3)-\psi^{\gamma_{t_k}}(p^a_1+p^a_2)) \\ &+ a^{\Psi,\Gamma}_{t_{k+1}}(Q)(\omega^a_4)(1-\psi^{\gamma_{t_k}}(p^a_1+p^a_2+p^a_3)). \end{aligned} \tag{3.17}$$

Continuing the backward recursion through the tree and applying the re-weighting of the reference probability measure according to the distortion function, we can calculate the bid and ask price of any non-negative future discounted cash flow at any point in time.

4 Application

To illustrate our methodology, we consider European options written on the S&P 500 index and we calibrate both a static and a stochastic liquidity model, as defined in Sections 3.3 and 3.3, to the bid-ask spreads of European index options.[15] All data comes from the OptionMetrics database accessed via the Wharton Research Data Services. For the calibration exercise, we choose as arbitrary date July 20, 2012. The S&P 500 had a closing price of $S_0 = 1362.66$ and the European options market was neither particularly stressed nor overly relaxed. At this day, the option data consists of 2560 calls and puts with maturities ranging from 7 days to 2.4 years and a strike interval of 100 to 3000. We removed 295 options from the dataset for whose mid price no Black-Scholes implied volatility could be calculated. For our discussion, we restrict our analysis to calls and puts within the $[80\%, 120\%]$ forward moneyness interval. Furthermore, we focus only on three maturities slices, namely on maturities of three, five, and eleven months. This leaves us with a total of 230 calls and puts.

4.1 Model specification

For our application, we assume that the log returns of the index are conditionally normal distributed.[16] For the liquidity process, we either assume it be constant for the static liquidity model or consider a mean-reverting square-root process following the findings of Albrecher et al. (2013) regarding the behavior of the static liquidity model parameter over time. In particular, for the stochastic liquidity model of Section 3.3, the asset price $S = (S_t)_{t \in \mathcal{T}_0^K}$ and the liquidity process $\Gamma = (\gamma_t)_{t \in \mathcal{T}_0^K}$ are binomial tree approximations of the continuous-time processes

$$dS_t = S_t(r-q)dt + S_t \sigma W_t^S \quad S_0 > 0, \tag{4.1}$$

$$d\gamma_t = \kappa(\theta - \gamma_t)dt + \nu\sqrt{\gamma_t} dW_t^\gamma, \quad \gamma_0 > 0, \tag{4.2}$$

[15] Using the same techniques, it is also possible to model bid and ask prices of path-dependent payoffs.

[16] This choice is merely for illustration purposes. We are well aware of the fact that there are more suitable choices for the underlying process. However, while, e.g., a stochastic volatility model such as the Heston model may be better suited, the imprecision in the volatility surface fit interferes with the assessment of the performance of the liquidity model.

where W^S and W^γ are correlated Brownian motions with $d\langle W^S, W^\gamma\rangle_t = \rho dt$, $\rho \in [-1, 1]$. By r we denote the one-period risk-free rate, q the dividend yield, and σ the asset volatility. To improve speed and memory usage, we prune the trees similar as in Baule and Wilkens (2004). For more details on the construction of the binomial trees, we refer to Appendix B. The asset price S in the static liquidity model is the same binomial tree approximation of (4.1) and the construction of the binomial tree follows similarly, except that γ is a constant now.

Following D. Madan et al. (2017), we modify the usual concave distortion function ψ^γ to take into account the step size h of the tree via

$$\psi^{\gamma,h}(u) := u + \sqrt{h}(\psi^\gamma(u) - u), \quad \forall u \in [0, 1], \tag{4.3}$$

which is still concave. Using this adjustment allows for an adequate comparison of parameter estimates for the market liquidity process for different step sizes and in particular the one-step static liquidity model. As distortion measure, we use CVaR.[17]

4.2 Calibration

In the following, we only refer to the calibration of the stochastic liquidity model. The calibration of the static model follows the same methodology and only differs in the constructed binomial tree (one-dimensional instead of two-dimensional) and number of liquidity parameters (one instead of five).[18]

Considering the available data in the different slices, it is evident that they are not equally distributed over either moneyness nor maturity. For example, the three month slices have over twice as many data points as the either of the other two and the eleven months moneyness interval [110%, 112%] contains over 20% of all data points despite being only 5% of the whole interval. To ensure a better fit over the whole moneyness range and not just areas with clustered strikes, we therefore calculate regularly interpolated bid and ask prices, P_i^{bid} and P_i^{ask} for $i \in \{1, ..., M\}$, for each slice.

[17] We also conducted our analysis by using different measures. The differences were insignificant.

[18] We remark that for our calibration exercise we follow the standard practice of calibrating, e.g., a stochastic volatility model, i.e., we treat the current level of liquidity as an additional parameter. A time-series estimation of implied liquidity would require setting up, e.g., a suitable filter method, which is beyond the scope of this paper.

These interpolated prices, instead of the real ones, will be used in our optimization algorithm below. The reported model fit in Table 1 and all figures will be based on real data.

We calibrate our model for each selected maturity slice of calls and puts separately. First, we calculate for each option $i \in \{1, ..., M\}$ the Black-Scholes implied volatility σ_i such that

$$\frac{1}{2}(P_i^{\text{ask}} + P_i^{\text{bid}}) = BS(S_0, K_i, T_i, r_i, q_i, \sigma_i), \tag{4.4}$$

where K_i and T_i are the strike and time-to-maturity of option i, respectively, r_i is the zero rate with maturity closest to T_i and q_i is the implied dividend rate from the put-call parity.[19] Furthermore, we denote the market bid and ask prices by P_i^{ask} and P_i^{bid}. For every option, the implied volatility parameter is then used to construct the binomial tree for the underlying together with the correlated tree for the liquidity process.[20] The risk-free rate, dividend yield and the liquidity parameters, including the correlation parameter ϱ, are the same over all options of the selected maturity slice while the implied volatility is allowed to be different to best replicate the observed market data.

Having fixed the distortion functions $\Psi = (\psi^z)_{z \geq 0}$ and a market liquidity process Γ in (4.2), which depends on the set of parameters $\Theta = \{\gamma_0, \kappa, \theta, \nu, \rho\}$, we can calibrate the model by minimizing the root-mean-square error (RMSE) of the normalized bid-ask spreads:[21]

$$\text{RMSE}(\Theta \,|\, \Psi, \Gamma) := \sqrt{\frac{1}{M} \sum_{i=1}^{M} \left(\frac{\Delta P_i^{\text{model}} - \Delta P_i^{\text{market}}}{\frac{1}{2}(P_i^{\text{bid}} + P_i^{\text{ask}})} \right)^2}, \tag{4.5}$$

where $\Delta P_i^{\text{market}}$ and $\Delta P_i^{\text{model}}$ are the market and model bid-ask spreads and M is the number of options. The model bid and ask prices are calculated by starting at the terminal cash flows and applying equations (3.16) respectively (3.17) backward recursively throughout the two-dimensional

[19] Inferring the dividend yield from the put-call parity partially circumvents problems caused by different quotation times from the option and underlying markets. The implied dividend yield for all options lies in the interval [1.90%, 2.21%]. The dividend yield reported in the OptionMetrics database is 2.50%.

[20] We remark that this procedure is only an approximation as the undistorted price is usually not exactly half-way between the bid and ask price in our model. The approximation becomes cruder when getting closer to maturity and further away from at the money. Nevertheless, compared to the alternative of having to calibrate M additional parameters, the error from this approach seems acceptable.

[21] We also tested minimizing the spreads in implied volatility, but found that the improved fit in the implied volatility space led to a, comparatively, bigger error in the normalized price space. Minimizing the model errors of the bid and ask prices respectively implied volatilities instead of the spreads did not lead to vastly different results.

binomial tree to get

$$\Delta P_i^{\text{model}} := a_0^{\Psi,\Gamma}(Q_i) - b_0^{\Psi,\Gamma}(Q_i), \tag{4.6}$$

where Q_i is the cash flow associated to the i^{th} option.[22]

To avoid the problem of local minima, we use a surrogate model based on radial basis functions.[23] Surrogate models are widely used in engineering because they require significantly less function evaluations than, e.g., genetic algorithms or particle swarm methods.

We first determined reasonable parameter ranges and optimized over these.[24] To help the numerical algorithm find better solutions, we then tightened the parameter ranges by determining where the best and worst model fits occurred. We tried to keep the parameter ranges as wide as possible to prevent influencing the final results too much.

4.3 Results and discussion

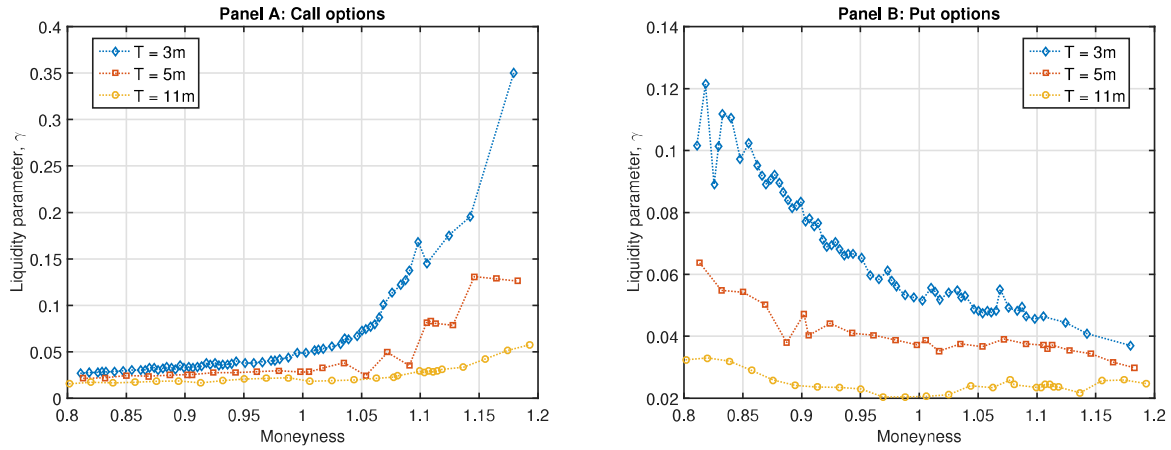


Figure 4. Market implied liquidity skew for selected maturity slices of the European options market on the S&P 500 on July 20, 2012. In total, we have 115 calls and the same number of puts. For every option, we calculate the static market liquidity parameter γ, which minimizes the bid-ask spread. We use the CVaR distortion function and assume a lognormal model for the underlying index.

To motivate the use of a stochastic liquidity component, we first consider a static liquidity model as in Section 3.3 for which we regard every option in isolation. For the CVaR distortion function,

[22] For the static liquidity model the relevant formulas are collected in Corollary 1.

[23] See Gutmann (2001) and in particular the toolbox MATSuMoTo developed by Mueller (2014).

[24] $\gamma \in [0.5\gamma^*, 2\gamma^*]$, $\kappa \in (0, 100]$, $\theta \in (0, 0.05]$, $\nu \in (0, 2]$, $\varrho \in (-1, 1)$. γ^* denotes the optimal liquidity parameter in the static model, see the next section for more details.

		Stochastic liquidity model						Static model	
Type	Mat	γ_0	κ	θ	ν	ρ	RMSE	γ^*	RMSE
call	3m	0.151	22.823	0.008	0.980	-0.951	9.4%	0.170	20.1%
call	5m	0.068	11.278	0.012	0.719	-0.974	3.4%	0.074	10.1%
call	11m	0.031	18.903	0.011	0.553	-0.983	1.4%	0.029	3.8%
put	3m	0.087	8.567	0.009	0.910	-0.975	1.2%	0.082	3.4%
put	5m	0.036	4.220	0.010	0.726	-0.952	0.7%	0.045	1.6%
put	11m	0.021	4.123	0.016	0.509	-0.977	0.6%	0.026	1.0%

Table 1. The table reports the calibrated parameters for European call and put options with different maturities (Mat) on the S&P 500 on July 20, 2012, for the stochastic liquidity model and the static model. The underlying distortion function is chosen to be CVaR. The parameters of the liquidity process are estimated by minimizing the RMSE of the normalized bid-ask spreads.

we calculate for every option $i \in \{1, ..., M\}$ the unique parameter γ_i such that the model bid-ask spreads calculated using Corollary 1 coincide with the market bid-ask spreads. In Figure 4, we plot the calibrated liquidity parameter γ for various maturity slices and levels of moneyness. Clearly, this market-implied liquidity parameter is far from being constant. It increases with decreasing maturity and when the option becomes out-of-the-money. Furthermore, the skew effect weakens with increasing maturity.[25]

Going one step forward, we calibrate our stochastic liquidity model given in equation (4.2). As in the static model, we calibrate each maturity slice separately. Table 1 reports the parameter estimates and the RMSEs for the calibration of three maturity slices. As in the static case, the current implied liquidity level γ_0 decreases with increasing maturity. Furthermore, liquidity shocks tend to be highly transitory, which is reflected by the high values for the estimates of the parameter κ. These values fit well our observation of the bid-ask spreads over time in Panel B of Figure 1. The persistence of liquidity shocks, however, tends to increase with increasing maturity. The volatility estimate ν decreases with increasing maturity. We find that for call options the long-term mean θ is almost at the same level for all three maturities. For puts, we find a similar pattern, except for the eleven months maturity slice. Finally, the estimates for the correlation parameter ρ indicate a consistent and rather strongly negative correlation between changes in price and liquidity. Such high values make intuitively sense, as in times of highly negative returns such as, e.g., during the global financial

[25]These properties of the implied liquidity parameter γ is in line with what has been observed in the static one-step models used in Albrecher et al. (2013) and Corcuera et al. (2012).

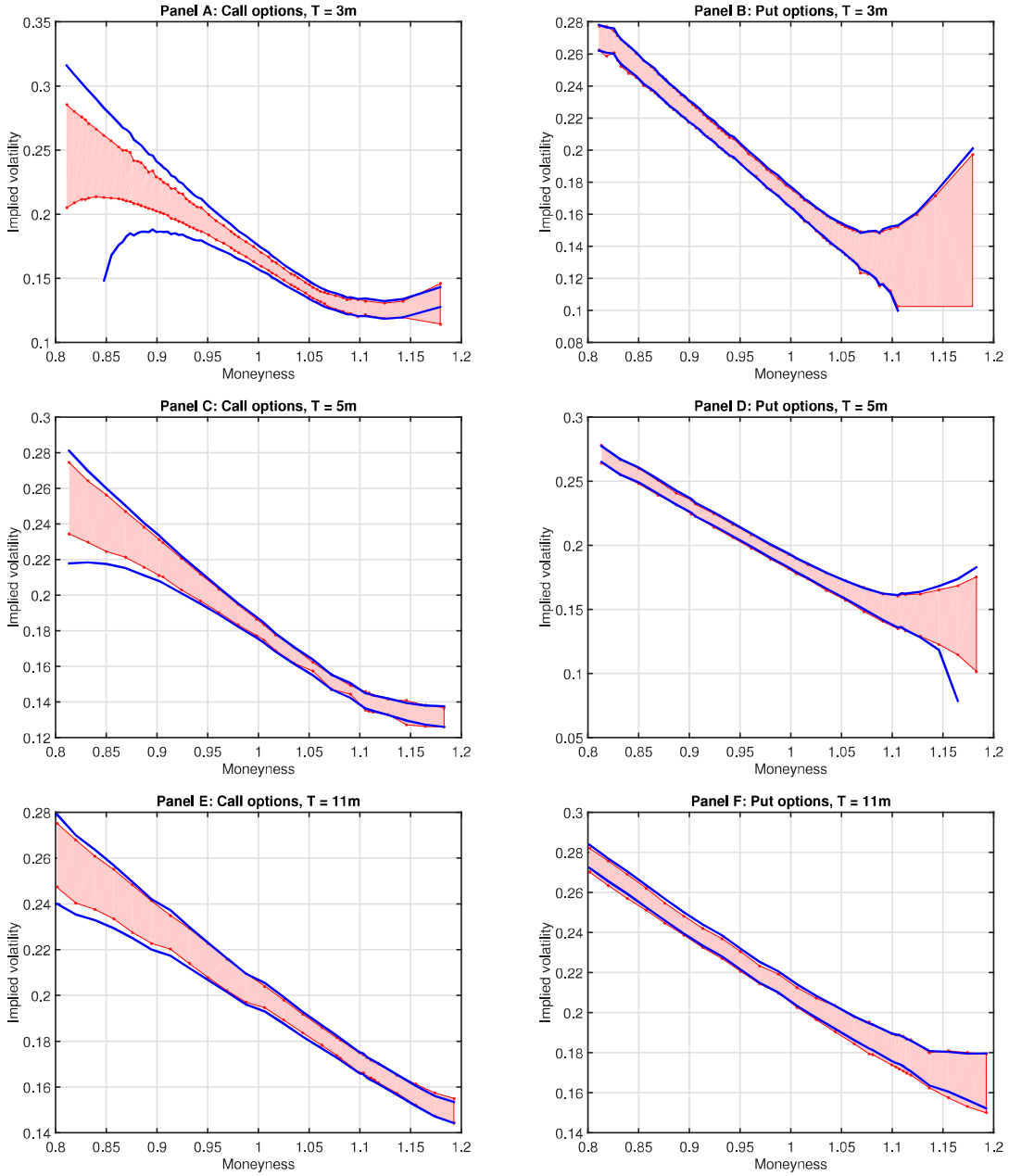


Figure 5. Calibration fit in the implied volatility space for European calls and puts on the S&P 500 on July 20, 2012. We plot the bid-ask spreads for maturities of 3, 5, and 11 months. We use the CVaR distortion function. The parameters of the liquidity process are estimated by minimizing the RMSE of the normalized bid-ask spreads. The shaded area corresponds to the market bid-ask spreads. The solid line corresponds to the model-implied spread.

crisis in 2008-2009 and the European crisis in 2012, bid-ask spreads also widened significantly.[26]

[26]The high correlation value corroborates the findings in Albrecher et al. (2013) of a high static liquidity parameter during times of crisis.

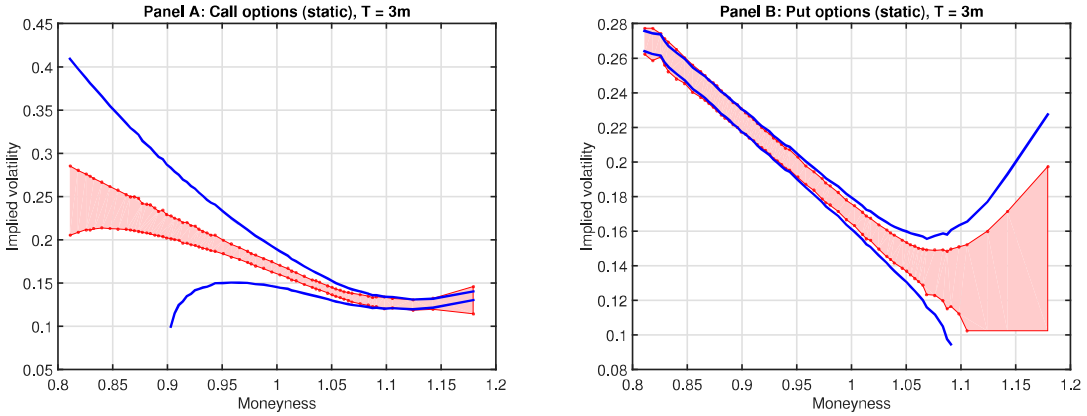


Figure 6. Calibration fit for the European call and put three months maturity slice on the S&P 500 on July 20, 2012. Panel A plots the bid-ask spread for calls, Panel B plots the spread in the implied volatility space for puts. The distortion function is chosen to be CVaR. The shaded area corresponds to the market bid-ask spreads. The solid line corresponds to the model-implied spread.

For comparison, we have also calibrated the static liquidity model for which all options on the selected maturity slices are fitted simultaneously. The optimal static liquidity parameter γ^* turns out to be of similar magnitude than the estimated γ_0 in the stochastic liquidity model. However, the model fit is considerably worse, with an RMSE often more than twice as large as the RMSE from the stochastic model. This leads us to the conclusion that the stochastic liquidity model can lead to significant improvements statistically, while also representing more closely stylized facts about market liquidity.[27]

To further illustrate our model's capability of fitting bid-ask spreads, we plot in Figure 5 the bid-ask prices of calls and puts in terms of implied volatilities. We first observe that the general behavior of the data is well replicated by the model for all slices and both option types. The model performs worse for shorter maturities, where spreads are also generally larger than for longer maturities. However, especially for short-term OTM put options, the bid-ask spread is fitted remarkably well. In contrast, as can be observed in Figure 6, the static model struggles to replicate the bid-ask spread at short maturities. Especially the errors for ITM options tend to be substantial.

[27] As an additional exercise, we have re-calibrated our model on October 10, 2008, in the wake of the financial crisis. Most parameters were all of the same order of magnitude. The calibration yielded parameters θ and ν that were higher than on July 20, 2012. Furthermore, the current level of liquidity γ_0 was also larger. We attribute this observation to the increased uncertainty in the market at the time and the much wider bid-ask spreads, especially for puts. The observations that κ is much higher for calls than puts and that ν decreases with maturity also held on October 10, 2008. Comparing with the static liquidity model, the RMSE was again improved by a factor of about two to three.

4.4 Parameter sensitivities

To gain more insights into the different roles of the parameters determining the liquidity process, we perform a sensitivity analysis for the put option with 5 months to maturity. Using the parameter values from Table 1, we assume different values for one parameter while keeping all others fixed. Figures 7 and 8 illustrate the sensitivities in the implied volatility and the normalized price space. Market implied bid-ask volatilities and prices correspond to the dashed lines marked with asterisks and are interpolated from S&P500 data on July 20, 2012.

Panels A and B of Figure 7 plot the impact of changes in the current liquidity level γ_0. As expected, high illiquidity, i.e., a large value for γ_0, is reflected by a wider bid-ask spread. For low values of γ_0, the spread collapses. In the implied volatility space, we note that the level of liquidity seems to impact options across all levels of moneyness in a similar way, i.e., by a parallel move of the implied volatility curve. Panels C and D plot the impact of changes in the parameter κ. Finally, in Panels E and F we plot the impact of changes in the parameter θ. Both parameters relate to the drift of the liquidity process. As for γ_0 in Panel A, we observe that changes in these parameters lead to a parallel move in the implied volatility surface. For more persistent liquidity shocks, i.e., a high value for κ, the bid-ask spread widens and it further does so when the long term mean for illiquidity θ is large.

In Figure 8 we plot the sensitivities with respect to the volatility of the liquidity process ν and the correlation between price changes and liquidity ρ. Here, we observe that the effect differs from the previous analysis in Figure 8 in that there is a less pronounced parallel impact on implied volatility. Instead, a change in the liquidity volatility parameter ν leads to a sharp increase in the convexity of the implied volatility of ask prices. Such an effect, but to a lesser extent, can also be observed for the implied correlation ρ.

5 Conclusion

Particularly after the financial crisis, the issue of market liquidity has been center stage for researchers and practitioners alike. However, the literature on liquidity in option markets and the discussion on

how to incorporate liquidity into the pricing problem has been sparse. We add to this literature by providing a theoretical framework that allows us to incorporate a stochastic liquidity process into the option pricing problem. By specifying a simple version of our model and taking it to the data, we find that the additional flexibility of having a stochastic liquidity component helps to replicate the bid and ask spreads typically observed in option markets.

In our empirical exercise, we focus on a very simple and illustrative example. Hence, it comes at no surprise that our calibration analysis indicates some challenging avenues for future research. For example, the lack of fit at short-maturities corroborates the need for adding a jump component to the liquidity process. Furthermore, it may be advantageous to model the underlying process using a stochastic volatility model which, of course, leads to challenging numerical problems. A natural candidate for such an extension could be the recombining stochastic volatility tree of Akyıldırım et al. (2014) as starting point. Inspired by the findings of Chou et al. (2011) it could be beneficial to consider a two-factor liquidity model, where one factor corresponds to the illiquidity of the underlying and the other to the illiquidity in the option market. Finally, one could go beyond a simple calibration exercise and try to perform a time-consistent estimation of the liquidity process using time series data. We leave these extensions to our analysis for future research.

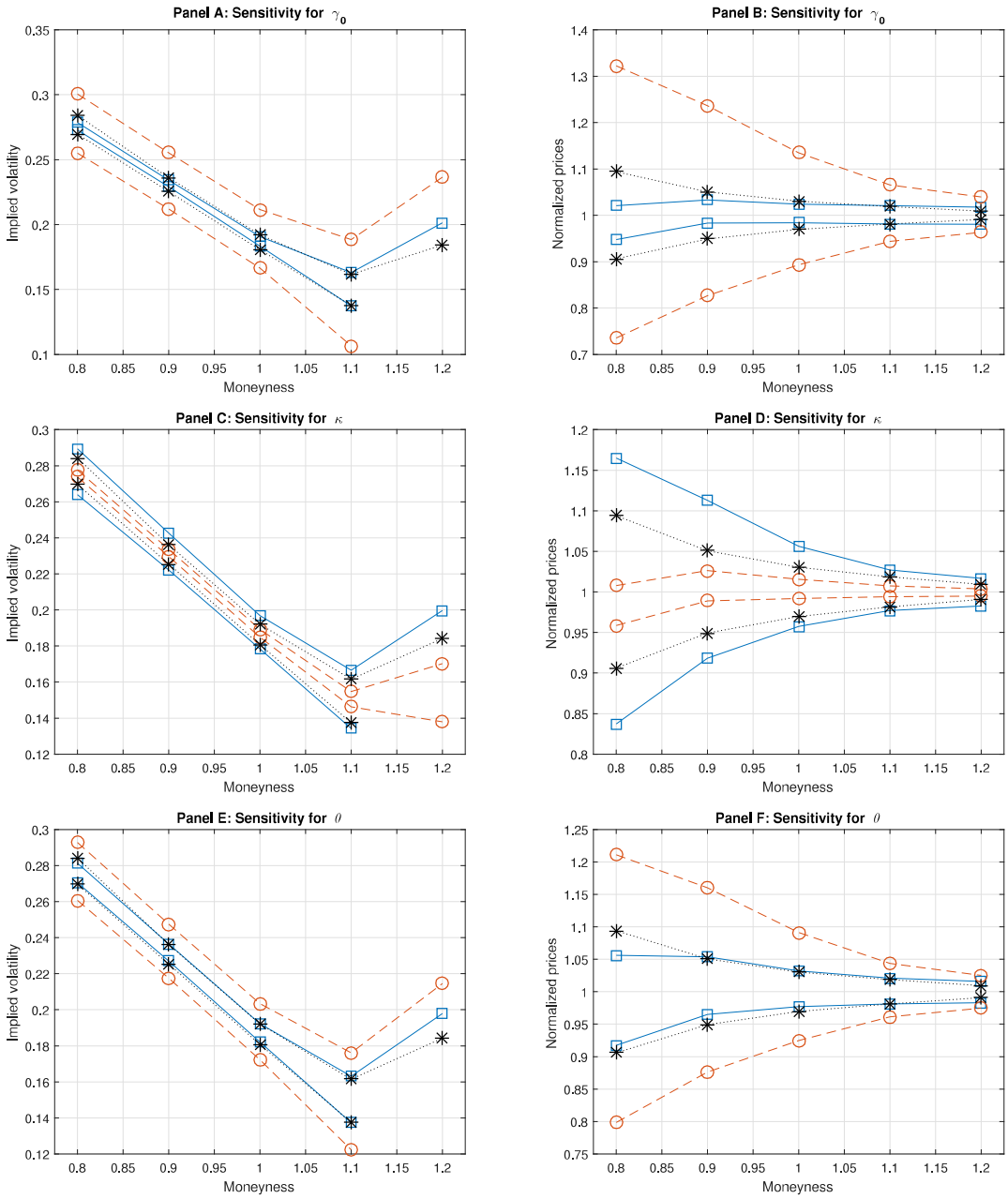


Figure 7. Sensitivity analysis for different parameter values. The figure plots the changes in the implied volatility and normalized price curve of the put option with 5 month maturity when we change the underlying liquidity parameters. Market implied bid-ask volatilities and prices correspond to the dashed lines marked with asterisks and are interpolated from S&P500 data on July 20, 2012. Panels A and B plot the impact of changes in γ_0. Solid (dashed) lines and squares (circles) correspond to $\gamma_0 = 0.01$ ($\gamma_0 = 0.2$). Panels C and D plot the impact of changes in the parameter κ. Solid (dashed) lines marked with squares (circles) correspond to $\kappa = 1$ ($\kappa = 30$). Panels E and F plot the impact of changes in the parameter θ. Solid (dashed) lines and squares (circles) correspond to $\theta = 0.001$ ($\theta = 0.1$). For all graphs, the remaining parameters were set equal to the values given in Table 1.

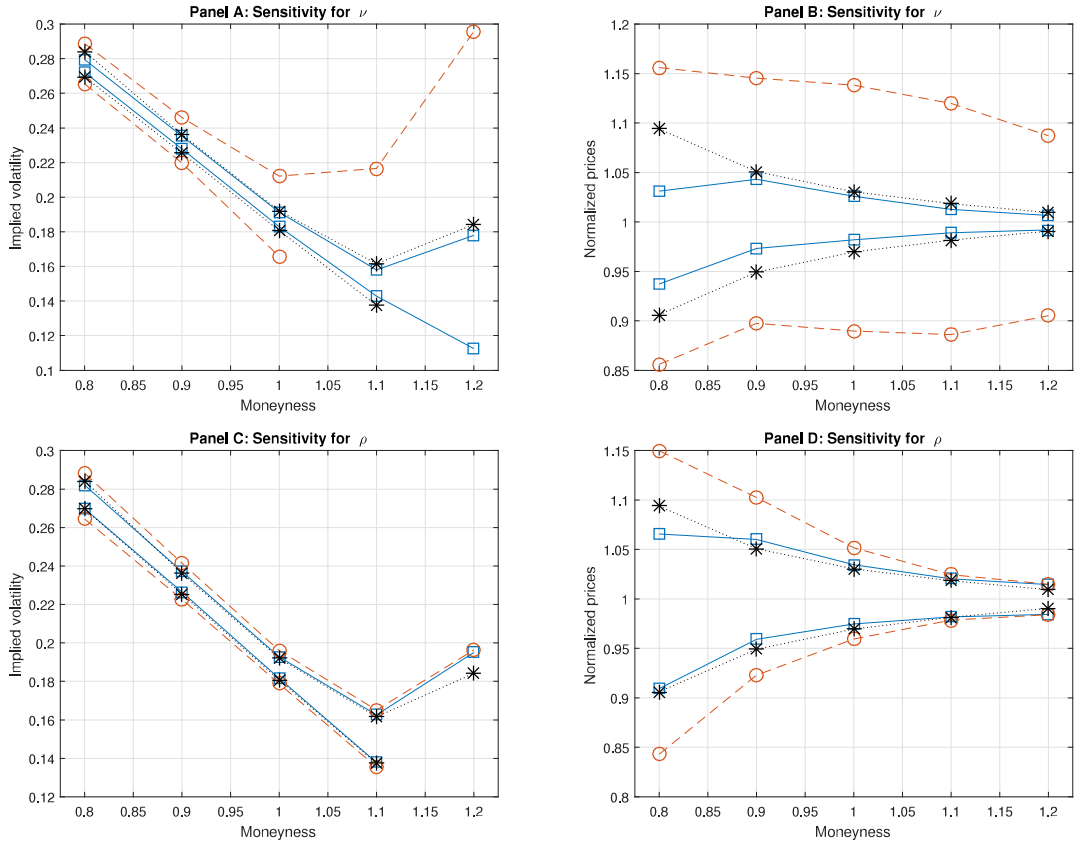


Figure 8. Sensitivity analysis for different parameter values. The figure plots the changes in the implied volatility and normalized price curve of the put option with 5 month maturity when we change the underlying liquidity parameters. Market implied bid-ask volatilities and prices correspond to the dashed lines marked with asterisks and are interpolated from S&P500 data on July 20, 2012. Panels A and B plot the impact of changes in the parameter ν. Solid (dashed) lines marked with squares (circles) correspond to $\nu = 0.25$ ($\nu = 0.15$). Panels C and D plot the impact of changes in the parameter ρ. Solid (dashed) lines and squares (circles) correspond to $\rho = -0.9$ ($\rho = 0.9$). For all graphs, the remaining parameters were set equal to the values given in Table 1.

References

Akyıldırım, E., Dolinsky, Y., & Soner, H. M. (2014). Approximating stochastic volatility by recombinant trees. *The Annals of Applied Probability*, *24*(5), 2176–2205.

Albrecher, H., Guillaume, F., & Schoutens, W. (2013). Implied liquidity: Model sensitivity. *Journal of Empirical Finance*, *23*, 48–67.

Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, *9*(3), 203–228.

Artzner, P., Delbaen, F., Eber, J.-M., Heath, D., & Ku, H. (2007). Coherent multiperiod risk adjusted values and Bellman's principle. *Annals of Operations Research*, *152*(1), 5–22.

Bannör, K. F., & Scherer, M. (2014). On the calibration of distortion risk measures to bid-ask prices. *Quantitative Finance*, *14*(7), 1217–1228.

Barles, G., & Soner, H. M. [Halil Mete]. (1998). Option pricing with transaction costs and a nonlinear Black-Scholes equation. *Finance and Stochastics*, *2*(4), 369–397.

Baule, R., & Wilkens, M. (2004). Lean trees - A general approach for improving performance of lattice models for option pricing. *Review of Derivatives Research*, *7*(1), 53–72.

Biagini, S., & Bion-Nadal, J. (2014). Dynamic quasi concave performance measures. *Journal of Mathematical Economics*, *55*, 143–153.

Bielecki, T. R., Cialenco, I., & Chen, T. (2015). Dynamic conic finance via backward stochastic difference equations. *SIAM Journal on Financial Mathematics*, *6*(1), 1068–1122.

Bielecki, T. R., Cialenco, I., Iyigunler, I., & Rodriguez, R. (2013). Dynamic conic finance: Pricing and hedging in market models with transaction costs via dynamic coherent acceptability indices. *International Journal of Theoretical and Applied Finance*, *16*(01), 1350002.

Bongaerts, D., De Jong, F., & Driessen, J. (2011). Derivative pricing with liquidity risk: Theory and evidence from the credit default swap market. *The Journal of Finance*, *66*(1), 203–240.

Çetin, U., Jarrow, R., Protter, P., & Warachka, M. (2006). Pricing options in an extended Black Scholes economy with illiquidity: Theory and empirical evidence. *Review of Financial Studies*, *19*(2), 493–529.

Çetin, U., Jarrow, R. A., & Protter, P. (2004). Liquidity risk and arbitrage pricing theory. *Finance and Stochastics*, *8*(3), 311–341.

Chan, K., & Chung, Y. P. (2012). Asymmetric price distribution and bid–ask quotes in the stock options market. *Asia-Pacific Journal of Financial Studies*, *41*(1), 87–102.

Cherny, A., & Madan, D. B. (2009). New measures for performance evaluation. *Review of Financial Studies*, *22*(7), 2571–2606.

Choquet, G. (1953). Theory of capacities, In *Annales de l'institut fourier*.

Chou, R. K., Chung, S.-L., Hsiao, Y.-J., & Wang, Y.-H. (2011). The impact of liquidity on option prices. *Journal of Futures Markets*, *31*(12), 1116–1141.

Christoffersen, P., Goyenko, R., Jacobs, K., & Karoui, M. (2018). Illiquidity premia in the equity options market. *The Review of Financial Studies*, *31*(3), 811–851.

Cohen, S. N., & Elliott, R. J. (2010a). Comparisons for backward stochastic differential equations on Markov chains and related no-arbitrage conditions. *The Annals of Applied Probability*, *20*(1), 267–311.

Cohen, S. N., & Elliott, R. J. (2010b). A general theory of finite state backward stochastic difference equations. *Stochastic Processes and their Applications*, *120*(4), 442–466.

Cohen, S. N., & Elliott, R. J. (2011). Backward stochastic difference equations and nearly time-consistent nonlinear expectations. *SIAM Journal on Control and Optimization*, *49*(1), 125–139.

Coquet, F., Hu, Y., Mémin, J., & Peng, S. (2002). Filtration-consistent nonlinear expectations and related g-expectations. *Probability Theory and Related Fields*, *123*(1), 1–27.

Corcuera, J. M., Guillaume, F., Madan, D. B., & Schoutens, W. (2012). Implied liquidity: Towards stochastic liquidity modelling and liquidity trading. *International Journal of Portfolio Analysis and Management*, *1*(1), 80–91.

Cvitanić, J., & Karatzas, I. (1996). Hedging and portfolio optimization under transaction costs: A martingale approach. *Mathematical Finance*, *6*(2), 133–165.

Davis, M. H., Panas, V. G., & Zariphopoulou, T. (1993). European option pricing with transaction costs. *SIAM Journal on Control and Optimization*, *31*(2), 470–493.

Denneberg, D. (1994). *Non-additive measure and integral*. Springer.

El Karoui, N., Peng, S., & Quenez, M. C. (1997). Backward stochastic differential equations in finance. *Mathematical Finance*, *7*(1), 1–71.

Engle, R., & Neri, B. (2010). The impact of hedging costs on the bid and ask spread in the options market. *Unpublished Working Paper, New York University.*

Ethier, S. N., & Kurtz, T. G. (2009). *Markov processes: Characterization and convergence.* John Wiley & Sons.

Feng, S.-P., Hung, M.-W., & Wang, Y.-H. (2014). Option pricing with stochastic liquidity risk: Theory and evidence. *Journal of Financial Markets*, *18*, 77–95.

Föllmer, H., & Schied, A. (2011). *Stochastic finance: An introduction in discrete time.* Walter de Gruyter.

George, T. J., & Longstaff, F. A. (1993). Bid-ask spreads and trading activity in the S&P 100 index options market. *Journal of Financial and Quantitative Analysis*, *28*(03), 381–397.

Gutmann, H.-M. (2001). A radial basis function method for global optimization. *Journal of Global Optimization*, *19*(3), 201–227.

Madan, D., Pistorius, M., & Stadje, M. (2017). On dynamic spectral risk measures, a limit theorem and optimal portfolio allocation. *Finance and Stochastics*, *21*(4), 1073–1102.

Madan, D. B. (2010). Conserving capital by adjusting deltas for gamma in the presence of skewness. *Journal of Risk and Financial Management*, *3*(1), 1–25.

Madan, D. B., & Cherny, A. (2010b). Markets as a counterparty: An introduction to conic finance. *International Journal of Theoretical and Applied Finance*, *13*(08), 1149–1177.

Madan, D. B., & Schoutens, W. (2014). Two processes for two prices. *International Journal of Theoretical and Applied Finance*, *17*(01), 1450005.

Mueller, J. (2014). MATSuMoTo: The MATLAB surrogate model toolbox for computationally expensive black-box global optimization problems. *arXiv preprint, arXiv:1404.4261.*

Pedersen, L. H. (2009). When everyone runs for the exit. *International Journal of Central Banking.*

Peng, S. (2007). G-expectation, G-Brownian motion and related stochastic calculus of Itô type. In *Stochastic analysis and applications* (pp. 541–567). Springer.

Riedel, F. (2004). Dynamic coherent risk measures. *Stochastic Processes and their Applications*, *112*(2), 185–200.

Rosazza Gianin, E., & Sgarra, C. (2013). Acceptability indexes via g-expectations: An application to liquidity risk. *Mathematics and Financial Economics*, *7*(4), 457–475.

Shreve, S. E., & Soner, H. M. [H Mete]. (1994). Optimal investment and consumption with transaction costs. *The Annals of Applied Probability*, *4*(3), 609–692.

Soner, H. M. [Halil M], Shreve, S. E., & Cvitanić, J. (1995). There is no nontrivial hedging portfolio for option pricing with transaction costs. *The Annals of Applied Probability*, *5*(2), 327–355.

Wang, S. S. (2000). A class of distortion operators for pricing financial and insurance risks. *Journal of Risk and Insurance*, *67*(1), 15–36.

Appendices

A Proof of Theorem 1

In the next two sections, we will summarize the theory of nonlinear expectations as solutions to BSΔEs on general discrete-time processes as developed by Cohen and Elliott (2010b). This framework is then used to show that the bid and ask prices from Definition 9 are in fact time-consistent and dynamically translation invariant nonlinear expectations as claimed in Theorem 1.

The log-price process of the underlying is denoted by $X = (X_t)_{t\in\mathcal{T}_0^K}$ which, like S, is a general discrete-time, finite state process. Cohen and Elliott (2010b) developed the theory for terminal conditions with multi-dimensional payoffs, but we will only present their results for the one-dimensional case. Additionally, while we assume without loss of generality that there are $N \in \mathbb{N}$ states at each time point $t \in \mathcal{T}_0^K$, it is worth noting that the theory can be extended to infinite states (see Cohen and Elliott (2011)), even though this is less relevant for our application which uses binomial trees.

A.1 BSΔE setup

Without loss of generality, we will set $X_0 = 0$ and assume that each X_{t_k}, $t_k \in \mathcal{T}_1^K$, takes values in the standard basis of $\mathbb{R}^N$, i.e.,

$$X_{t_k} \in \{e_1, \cdots, e_N\}, \quad e_j := (0, \cdots, 0, 1, 0, \cdots, 0)^\top \in \mathbb{R}^N, \tag{A.1}$$

where e_j is one at the j-th component and $(\cdot)^\top$ denotes the transposition operator. We call the process $M := (M_{t_k})_{k=0}^K$ defined by

$$M_{t_k} := X_{t_k} - \mathbb{E}[X_{t_k} \,|\, \mathcal{F}_{t_{k-1}}] \in \mathbb{R}^N, \quad k = 1, ..., K \tag{A.2}$$

and $M_0 := 0$ the martingale difference process.

Definition 10 (BSΔE)**.** *Let $(Y, Z) := (Y_t, Z_t)_{t\in\mathcal{T}_0^K}$ be $\mathbb{R} \times \mathbb{R}^N$-valued adapted processes, $F : \Omega \times \mathcal{T}_0^K \times \mathbb{R} \times \mathbb{R}^N \longrightarrow \mathbb{R}$ an adapted function and Q an $\mathbb{R}$-valued, $\mathcal{F}_T$-measurable random variable.*

We say (Y, Z) is a solution of the BSΔE based on M with driver F and terminal condition Q if

and only if (Y, Z) *satisfies for all* $\omega \in \Omega$ *and* $t_k \in \mathcal{T}_0^K$,

$$Y_{t_k}(\omega) - \sum_{t_i \in \mathcal{T}_k^{K-1}} F(\omega, t_i, Y_{t_i}(\omega), Z_{t_i}(\omega)) + \sum_{t_i \in \mathcal{T}_k^{K-1}} Z_{t_i}(\omega) M_{t_{i+1}}(\omega) = Q(\omega). \tag{A.3}$$

From now on we will omit the argument $\omega \in \Omega$ of M, Q, X, Y and Z. Also note that the BSΔE can be equivalently written in difference form as

$$\begin{aligned} Y_{t_k} - F(\omega, t_k, Y_{t_k}, Z_{t_k}) + Z_{t_k} M_{t_{k+1}} = Y_{t_{k+1}} \quad \forall t_k \in \mathcal{T}_0^{K-1} \\ Y_T = Q. \end{aligned} \tag{A.4}$$

A BSΔE is the discrete-time version of a BSDE (see El Karoui et al. (1997)), i.e., for all $t \in [0, T]$ and appropriately defined functions and variables

$$Y_t - \int_t^T F(\omega, u, Y_{u-}, Z_u) du + \int_t^T Z_u dM_u = Q. \tag{A.5}$$

Even though the solution to the BSΔE is given by the pair (Y, Z), we are not particularly interested in Z and will only implicitly use it. Indeed, subtracting from the BSΔE its conditional expectation shows that the process $(Z_{t_k} M_{t_{k+1}})_{t_k \in \mathcal{T}_0^{K-1}}$ can be expressed only in terms of $(Y_{t_{k+1}})_{t_k \in \mathcal{T}_0^{K-1}}$.

Lemma 1. *If* (Y, Z) *are adapted solutions to the BSΔE, then it holds for* $t_k \in \mathcal{T}_0^{K-1}$

$$Z_{t_k} M_{t_{k+1}} = Y_{t_{k+1}} - \mathbb{E}[Y_{t_{k+1}} \,|\, \mathcal{F}_{t_k}]. \tag{A.6}$$

Proof. The BSΔE is given by, for $t_k \in \mathcal{T}_0^{K-1}$,

$$Y_{t_k} - F(\omega, t_k, Y_{t_k} Z_{t_k}) + Z_t M_{t_{k+1}} = Y_{t_{k+1}}. \tag{A.7}$$

Taking the conditional expectation on both sides gives

$$Y_{t_k} - F(\omega, t_k, Y_{t_k} Z_{t_k}) = \mathbb{E}[Y_{t_{k+1}} \,|\, \mathcal{F}_{t_k}], \tag{A.8}$$

since Y, Z and F are adapted and $\mathbb{E}[M_{t_{k+1}} \,|\, \mathcal{F}_{t_k}] = 0$. Subtracting these two equations gives the result. □

We will in the following assume that all solutions (Y, Z) satisfy

$$Y_t, Z_t \in L^1(\mathcal{F}_t) \quad \forall t \in \mathcal{T}_0^K, \tag{A.9}$$

that the terminal condition Q is in $L^1(\mathcal{F}_T)$ and that the driver fulfills

$$F(\omega, t, Y_t, Z_t) \in L^1(\mathcal{F}_t) \quad \forall \omega \in \Omega, t \in \mathcal{T}_0^K, Y_t, Z_t \in L^1(\mathcal{F}_t). \tag{A.10}$$

Definition 11 (Equivalent processes)**.** *Let Z^1, Z^2 be two $\mathbb{R}^N$-valued adapted process. We call Z^1 and Z^2 equivalent at time $t_k \in \mathcal{T}_0^{K-1}$, denoted as $Z^1_{t_k} \sim_{M_{t_{k+1}}} Z^2_{t_k}$, if and only if*

$$Z^1_{t_k} M_{t_{k+1}} = Z^2_{t_k} M_{t_{k+1}} \quad \mathbb{P}\text{-a.s.} \tag{A.11}$$

Z^1 and Z^2 are equivalent and we write $Z^1 \sim_M Z^2$ if and only if Z^1 and Z^2 are equivalent at all times in $\mathcal{T}_0^{K-1}$.

Under suitable assumptions, it is now possible to show that each BSΔE has a solution that is unique up to equivalence.

Theorem 2. *Assume the driver $F : \Omega \times \mathcal{T}_0^K \times \mathbb{R} \times \mathbb{R}^N \longrightarrow \mathbb{R}$ satisfies*

(i) for all $\mathbb{R}$-valued adapted processes Y, $\mathbb{R}^N$-valued adapted processes Z^1, Z^2, $t \in \mathcal{T}_0^K$ and $\mathbb{P}$-almost all $\omega \in \Omega$, if $Z^1 \sim_M Z^2$ then

$$F(\omega, t, Y_t, Z^1_t) = F(\omega, t, Y_t, Z^2_t). \tag{A.12}$$

(ii) for all $z \in \mathbb{R}^N$, $t \in \mathcal{T}_0^K$ and $\mathbb{P}$-almost all $\omega \in \Omega$, the map $y \longmapsto y - F(\omega, t, y, z)$ is a bijection from $\mathbb{R} \longrightarrow \mathbb{R}$.

Then, for all $Q \in L^1(\mathcal{F}_T)$, the BSΔE (A.3) has an adapted and $\mathbb{R} \times \mathbb{R}^N$-valued solution (Y, Z) that is unique up to indistinguishability for Y and $\sim_M$ for Z.

Proof. Cohen and Elliott (2010b), Theorem 2. □

Definition 12 (Set of indices). *The $\mathcal{F}_{t_k}$-measurable set of indices of possible values of $X_{t_{k+1}}$ given $\mathcal{F}_{t_k}$ is defined as*

$$\mathbb{J}_{t_k} := \{j \in \{1, ..., N\} \,|\, \mathbb{P}(X_{t_{k+1}} = e_j \,|\, \mathcal{F}_{t_k}) > 0\}, \tag{A.13}$$

where the e_j are the standard basis of $\mathbb{R}^N$ as before and $t_k \in \mathcal{T}_0^{K-1}$.

A.2 Connection between BSΔEs and nonlinear expectations

Definition 13 (Balanced driver). *A driver F is called balanced if and only if it satisfies assumptions (i) and (ii) of Theorem 2 and furthermore, for all $Q^1, Q^2 \in L^1(\mathcal{F}_T)$, the corresponding BSΔE solutions (Y^1, Z^1), (Y^2, Z^2) satisfy*

(iii) for all $t_k \in \mathcal{T}_0^K$ and $\mathbb{P}$-almost all $\omega \in \Omega$,

$$F^1(\omega, t_k, Y^2_{t_k}, Z^1_{t_k}) - F^1(\omega, t_k, Y^2_{t_k}, Z^2_{t_k}) \geq \min_{j \in \mathbb{J}_{t_k}} \{(Z^1_{t_k} - Z^2_{t_k})(e_j - \mathbb{E}[X_{t_{k+1}} \,|\, \mathcal{F}_{t_k}]\}. \tag{A.14}$$

and equality holds only if $Z^1_{t_k} \sim_{M_{t_{k+1}}} Z^2_{t_k}$.

(iv) if for all $t \in \mathcal{T}_0^K$ and $\mathbb{P}$-almost all $\omega \in \Omega$

$$Y^1_t - F(\omega, t, Y^1_t, Z^1_t) \geq Y^2_t - F(\omega, t, Y^2_t, Z^2_t), \tag{A.15}$$

then $Y^1_t \geq Y^2_t$ $\mathbb{P}$-a.s.

Definition 14 (Normalized driver). *A driver F is called normalized if and only if for all $t \in \mathcal{T}_0^K$, $\mathbb{R}$-valued, $\mathcal{F}_t$-measurable processes Y and $\mathbb{P}$-almost all $\omega \in \Omega$, it holds that $F(\omega, t, Y, 0) = 0$.*

This brings us to the main theorem which gives a one-to-one connection between time-consistent nonlinear expectations and solutions to BSΔEs.

Theorem 3. *Let $(\mathcal{E}(\cdot \,|\, \mathcal{F}_t))_{t \in \mathcal{T}_0^K}$ an time-consistent nonlinear expectation.*

Then, the following are equivalent.

(i) $(\mathcal{E}(\cdot \,|\, \mathcal{F}_t))_{t \in \mathcal{T}_0^K}$ is dynamically translation invariant.

(ii) There exists a driver F that is balanced, independent of Y and normalized such that for each $Q \in L^1(\mathcal{F}_T)$, $Y_t = \mathcal{E}(Q \,|\, \mathcal{F}_t)$ is a solution to the BSΔE (A.3) with terminal condition Q and driver F.

Furthermore, the two statements are connected via

$$F(\omega, t_k, Y_{t_k}, Z_{t_k}) = \mathcal{E}(Z_{t_k} M_{t_{k+1}} \,|\, \mathcal{F}_{t_k}) \quad \forall t_k \in \mathcal{T}_0^K. \tag{A.16}$$

Proof. Cohen and Elliott (2010b), Theorem 7. □

We will exploit this result by defining a driver F according to our one-period intuition that satisfies the assumptions of Theorem 3, (ii). Then, by taking the conditional expectation of the BSΔE (A.4), we get the backward recursive formula

$$\begin{aligned} \mathcal{E}(Q \,|\, \mathcal{F}_{t_k}) = Y_{t_k} &= \mathbb{E}[Y_{t_{k+1}} \,|\, \mathcal{F}_{t_k}] + F(\omega, t_k, Y_{t_k}, Z_{t_k}) \quad \forall t_k \in \mathcal{T}_0^{K-1} \\ Y_T &= Q, \end{aligned} \tag{A.17}$$

which allows us to calculate the time-consistent nonlinear expectation operator based on such a driver.

Remark 1. *In the static model, coherent (distortion) risk measures are used to define nonlinear expectations. In this discrete-time extension it is the other way around as defining*

$$\varrho_t(Q) := -\mathcal{E}(Q \,|\, \mathcal{F}_t) \quad \forall t \in \mathcal{T}_0^K \tag{A.18}$$

gives a dynamic risk measure. This was proposed by Rosazza Gianin and Sgarra (2013) in a continuous-time Brownian Motion setting. Cohen and Elliott (2010a) prove that $(\varrho_t)_{t \in \mathcal{T}_0^K}$ satisfy all the necessary properties, provided the driver F fulfills the assumptions of Theorem 3, (ii). In particular, translation invariance follows by normalization and independence of Y and monotonicity by F being balanced. Positive homogeneity requires F to be positively homogeneous itself. Furthermore, if F is convex, $(\varrho_t)_{t \in \mathcal{T}_0^K}$ is subadditive and hence coherent.

A.3 The bid and ask drivers

In general, $(\mathbb{E}_t^\psi[\cdot])_{t\in\mathcal{T}_0^K}$ is no time-consistent nonlinear and dynamically translation invariant expectation, though it still satisfies some of the properties as the following Lemma shows.

Lemma 2. *Let $(\mathbb{E}_t^\psi[\cdot])_{t\in\mathcal{T}_0^K}$ be a distorted conditional expectation. Then, for all $t \in \mathcal{T}_0^K$, $\mathbb{E}_t^\psi[\cdot]$ satisfies for all $Q^1, Q^2 \in L^1(\mathcal{F}_T)$ and $q \in L^1(\mathcal{F}_t)$*

(i) adaptedness, i.e., $\mathbb{E}_t^\psi[q] = q$ $\mathbb{P}$-a.s.

(ii) monotonicity, i.e., if $Q^1 \leq Q^2$ $\mathbb{P}$-a.s. then $\mathbb{E}_t^\psi[Q^1] \leq \mathbb{E}_t^\psi[Q^2]$ $\mathbb{P}$-a.s. and in particular, equality holds if and only if $Q^1 = Q^2$ $\mathbb{P}$-a.s.

(iii) dynamic translation invariance, i.e., $\mathbb{E}_t^\psi[Q+q] = \mathbb{E}_t^\psi[Q] + q$ $\mathbb{P}$-a.s.

Proof. Let $t \in \mathcal{T}$.

(i) Let $q \in L^1(\mathcal{F}_t)$. Now, since q is $\mathcal{F}_t$-measurable, $\mathbb{1}_{\{q\leq x\}}$ is as well for all $x \in \mathbb{R}$ and hence,

$$\mathbb{P}(q \leq x \,|\, \mathcal{F}_t) = \mathbb{E}[\mathbb{1}_{\{q\leq x\}} \,|\, \mathcal{F}_t] = \mathbb{1}_{\{q\leq x\}} \quad \forall x \in \mathbb{R}. \tag{A.19}$$

This, combined with the fact that

$$\psi(\omega, \mathbb{1}_A) = \mathbb{1}_A \quad \forall \omega \in \Omega \tag{A.20}$$

for every set $A \in \mathcal{F}_T$, directly gives, for each $\omega \in \Omega$,

$$\begin{aligned}
\mathbb{E}_t^\psi[q](\omega) &= -\int_{-\infty}^0 \psi(\omega, \mathbb{1}_{\{q\leq x\}}(\omega))dx + \int_0^\infty 1 - \psi(\omega, \mathbb{1}_{\{q\leq x\}}(\omega))dx \\
&= -\int_{-\infty}^0 \mathbb{1}_{\{q\leq x\}}(\omega)dx + \int_0^\infty 1 - \mathbb{1}_{\{q\leq x\}}(\omega)dx \\
&= -\int_{-\infty}^0 \mathbb{1}_{\{q\leq x\}}(\omega)dx + \int_0^\infty \mathbb{1}_{\{q>x\}}(\omega)dx \\
&= q(\omega).
\end{aligned} \tag{A.21}$$

(ii) The monotonicity property of $\mathbb{E}_t^\psi[\cdot](\omega)$ for all $\omega \in \Omega$ follows by observing that the conditional probability measure in a single state, $P_t(\cdot)(\omega)$, is a probability measure. Therefore,

$\psi(\omega, P_t(\cdot)(\omega))$ is a finite monotone set function and hence the Choquet integral defined thereby is monotone.

(iii) Let $Q \in L^1(\mathcal{F}_T)$, $q \in L^1(\mathcal{F}_t)$, $x \in \mathbb{R}$ and $\omega \in \Omega$. Since q is $\mathcal{F}_t$-measurable, the conditional probability on $\mathcal{F}_t$ at ω of $\{Q + q \leq x\}$ only depends on $q(\omega)$, i.e.,

$$\mathbb{P}(Q + q \leq x \,|\, \mathcal{F}_t)(\omega) = \mathbb{P}(Q + q(\omega) \leq x \,|\, \mathcal{F}_t)(\omega). \tag{A.22}$$

We use this and the change of measure $y := x - q(\omega)$ to get

$$\begin{aligned}
\mathbb{E}_t^{\psi}[Q+q](\omega) &= -\int_{-\infty}^{0} \psi(\omega, \mathbb{P}(Q \leq x - q(\omega) \,|\, \mathcal{F}_t)(\omega))dx \\
&\quad + \int_{0}^{\infty} 1 - \psi(\omega, \mathbb{P}(Q \leq x - q(\omega) \,|\, \mathcal{F}_t)(\omega))dx \\
&= -\int_{-\infty}^{-q(\omega)} \psi(\omega, \mathbb{P}(Q \leq y \,|\, \mathcal{F}_t)(\omega))dy \\
&\quad + \int_{-q(\omega)}^{\infty} 1 - \psi(\omega, \mathbb{P}(Q \leq y \,|\, \mathcal{F}_t)(\omega))dy \\
&= \mathbb{E}_t^{\psi}[Q](\omega) + q(\omega).
\end{aligned} \tag{A.23}$$

□

Corollary 2. *$(-\mathbb{E}_t^{\psi}[-\cdot])_{t \in \mathcal{T}_0^K}$ satisfies the same properties.*

Proof. Follows directly from Lemma 2. □

We now define the drivers for the bid and ask price

Definition 15 (Bid and ask driver)**.** *Let $\Psi = (\psi^z)_{z \geq 0}$ be a family of concave distortion functions that are pointwise increasing in z. For the market liquidity process $\Gamma = (\gamma_t)_{t \in \mathcal{T}_0^K}$, we define the bid driver to be the function given by*

$$F^{b,\Psi,\Gamma}(\omega, t_k, Y_{t_k}, Z_{t_k}) := \mathbb{E}_{t_k}^{\psi^{\gamma_{t_k}}}[Z_{t_k} M_{t_{k+1}}] \tag{A.24}$$

and the ask driver is set to be

$$F^{a,\Psi,\Gamma}(\omega, t_k, Y_{t_k}, Z_{t_k}) := -\mathbb{E}_{t_k}^{\psi^{\gamma_{t_k}}}[-Z_{t_k} M_{t_{k+1}}] \tag{A.25}$$

for all $\omega \in \Omega$, $t_k \in \mathcal{T}_0^K$, (Y, Z) $\mathbb{R} \times \mathbb{R}^N$*-valued adapted processes.*

By the definition of Choquet integrals and because we only consider concave distortion functions, the bid driver is convex and the ask driver is concave. Additionally, both are positively homogeneous and satisfy the following other properties.

Lemma 3. *The drivers* $F^{b,\Psi,\Gamma}$ *and* $F^{a,\Psi,\Gamma}$ *satisfy the assumptions of Theorem 3 (ii), i.e.,*

(i) independence of Y*.*

(ii) balanced on $L^1(\mathcal{F}_T)$*.*

(iii) normalization.

(iv) for each $Q \in L^1(\mathcal{F}_T)$*, there exists a solution to the BS*Δ*E (A.3) with terminal condition* Q *and driver* $F^{b,\Psi,\Gamma}$ *respectively* $F^{a,\Psi,\Gamma}$*.*

Proof. Denote $F^{b,\Psi,\Gamma}$ and $F^{a,\Psi,\Gamma}$ both by F.

(i) Holds by definition.

(ii) Due to the independence of Y and since $F(\omega, t_k, Y_{t_k}, Z^1_{t_k}) = F(\omega, t_k, Y_{t_k}, Z^2_{t_k})$ $\mathbb{P}$-a.s. if $Z^1_{t_k} \sim_{M_{t_{k+1}}} Z^2_{t_k}$ by definition, the only difficulty lies in proving that for all $\omega \in \Omega$ and $t_k \in \mathcal{T}_0^K$,

(a) for all $Q^1, Q^2 \in L^1(\mathcal{F}_T)$ and their corresponding solutions (Y^1, Z^1) and (Y^2, Z^2) it holds that $\mathbb{P}$-a.s.

$$F(\omega, t_k, Y^2_{t_k}, Z^1_{t_k}) - F(\omega, t_k, Y^2_{t_k}, Z^2_{t_k}) \geq \min_{j \in \mathbb{J}_{t_k}} \{(Z^1_{t_k} - Z^2_{t_k})(e_j - \mathbb{E}[X_{t_{k+1}} \,|\, \mathcal{F}_{t_k}])\}. \tag{A.26}$$

(b) equality holds only if $Z^1_{t_k} \sim_{M_{t_{k+1}}} Z^2_{t_k}$.

The proof turns out to be essentially the same as the one of Theorem 3 given by Cohen and Elliott (2010b), where the driver F satisfies the necessary properties of $(\mathcal{E}(\cdot \,|\, \mathcal{F}_{t_k}))_{t_k \in \mathcal{T}_0^K}$. Recall that

$$\mathbb{J}_{t_k} = \{i \in \{1, ..., N\} \,|\, \mathbb{P}(X_{t_{k+1}} = e_i \,|\, \mathcal{F}_{t_k}) > 0\}. \tag{A.27}$$

(a) Define a $\mathcal{F}_{t_k}$-measurable random variable q by

$$q := \min_{j \in \mathbb{J}_{t_k}} \{Y^1_{t_{k+1}} - Y^2_{t_{k+1}} \,|\, \mathcal{F}_{t_k}, X_{t_{k+1}} = e_j\}. \tag{A.28}$$

By Lemma 1 we know that

$$Z_{t_k} M_{t_{k+1}} = Y_{t_{k+1}} - \mathbb{E}[Y_{t_{k+1}} \,|\, \mathcal{F}_{t_k}], \tag{A.29}$$

where Y and Z denote Y^1 and Z^1 respectively Y^2 and Z^2. Plugging this into the definition of the random variable q gives

$$q = \mathbb{E}[Y^1_{t_{k+1}} \,|\, \mathcal{F}_{t_k}] - \mathbb{E}[Y^2_{t_{k+1}} \,|\, \mathcal{F}_{t_k}] + (Z^1_{t_k} - Z^2_{t_k}) \min_{j \in \mathbb{J}_{t_k}} \{e_j - \mathbb{E}[X_{t_{k+1}} \,|\, \mathcal{F}_{t_k}]\}, \tag{A.30}$$

where we used the definition of M as

$$M_{t_{k+1}} := X_{t_{k+1}} - \mathbb{E}[X_{t_{k+1}} \,|\, \mathcal{F}_{t_k}]. \tag{A.31}$$

By the definition of q, it holds that $\mathbb{P}$-a.s.

$$Y^1_{t_{k+1}} - q \geq Y^2_{t_{k+1}} \tag{A.32}$$

and since $\mathbb{E}^\psi$ is monotone by Lemma 2, we have $\mathbb{P}$-a.s.

$$\mathbb{E}^\psi_{t_k}[Y^1_{t_{k+1}} - q] \geq \mathbb{E}^\psi_{t_k}[Y^2_{t_{k+1}}]. \tag{A.33}$$

However, since $\mathbb{E}^\psi$ is also dynamically translation invariant by the same Lemma, this is equivalent to

$$\mathbb{E}^\psi_{t_k}[Y^1_{t_{k+1}}] - \mathbb{E}^\psi_{t_k}[Y^2_{t_{k+1}}] \geq q. \tag{A.34}$$

Subtracting $\mathbb{E}[Y^1_{t_{k+1}} - Y^2_{t_{k+1}} \,|\, \mathcal{F}_{t_k}]$ from both sides and using the definition of F and q as well as again the dynamic translation invariance of $\mathbb{E}^\psi$ gives the claim.

(b) We use the same random variable q as before and additionally assume equality which by

the previous argument is equivalent to assuming that

$$q = \mathbb{E}^{\psi}_{t_k}[Y^1_{t_{k+1}}] - \mathbb{E}^{\psi}_{t_k}[Y^2_{t_{k+1}}]. \tag{A.35}$$

Comparing this with expression (A.30) shows that

$$q = Y^1_{t_{k+1}} - Y^2_{t_{k+1}}. \tag{A.36}$$

Therefore,

$$Y^1_{t_{k+1}} - Y^2_{t_{k+1}} = \mathbb{E}^{\psi}_{t_k}[Y^1_{t_{k+1}}] - \mathbb{E}^{\psi}_{t_k}[Y^2_{t_{k+1}}] \tag{A.37}$$

which is equivalent to

$$Z^1_{t_k} \sim_{M_{t_{k+1}}} Z^2_{t_k}. \tag{A.38}$$

(iii) Holds by definition.

(iv) Holds by Theorem 2.

□

Remark 2. *In the following we will demonstrate that D. B. Madan's 2010 proof of showing that the drivers are balanced is incorrect. Let us rewrite Madan's argument using our notation.*

Proof in D. B. Madan (2010). Recall that,

$$M_{t_{k+1}} = X_{t_{k+1}} - \mathbb{E}[X_{t_{k+1}} \,|\, \mathcal{F}_{t_k}]. \tag{A.39}$$

Then, by the definition of the bid driver as the distorted conditional expectation of $Z_{t_k} M_{t_{k+1}}$ and because $X_{t_{k+1}}$ takes values in the standard basis of $\mathbb{R}^N$,

$$F^b(\omega, t_k, Y^2_{t_k}, Z^1_{t_k}) \geq \min_{j \in \mathbb{J}_{t_k}} \{Z^1_{t_k}(e_j - \mathbb{E}[X_{t_{k+1}} \,|\, \mathcal{F}_{t_k}]\} \tag{A.40}$$

and

$$F^b(\omega, t_k, Y^2_{t_k}, Z^2_{t_k}) \leq \max_{j \in \mathbb{J}_{t_k}} \{Z^2_{t_k}(e_j - \mathbb{E}[X_{t_{k+1}} \,|\, \mathcal{F}_{t_k}]\}. \tag{A.41}$$

Therefore,

$$F(\omega, t_k, Y^2_{t_k}, Z^1_{t_k}) - F(\omega, t_k, Y^2_{t_k}, Z^2_{t_k}) \geq \min_{j \in \mathbb{J}_{t_k}} \{(Z^1_{t_k} - Z^2_{t_k})(e_j - \mathbb{E}[X_{t_{k+1}} \,|\, \mathcal{F}_{t_k}]\}. \tag{A.42}$$

□

The above line of reasoning is clearly wrong. What the argument used in the above proof of D. B. Madan (2010) actually shows is that

$$\begin{aligned} &F(\omega, t_k, Y^2_{t_k}, Z^1_{t_k}) - F(\omega, t_k, Y^2_{t_k}, Z^2_{t_k}) \\ &\qquad \geq \min_{j \in \mathbb{J}_{t_k}} \{Z^1_{t_k}(e_j - \mathbb{E}[X_{t_{k+1}} \,|\, \mathcal{F}_{t_k}]\} + \min_{j \in \mathbb{J}_{t_k}} \{-Z^2_{t_k}(e_j - \mathbb{E}[X_{t_{k+1}} \,|\, \mathcal{F}_{t_k}]\}. \end{aligned} \tag{A.43}$$

The minimizing indices j_1^* of equation (A.40) and j_2^* of equation (A.41) do not in general coincide with the minimizing index j^* of equation (A.42). Therefore, the lower bound of $F(\omega, t_k, Y^2_{t_k}, Z^1_{t_k}) - F(\omega, t_k, Y^2_{t_k}, Z^2_{t_k})$ that needs to hold for the balanced property, given in equation (A.42), is always greater or equal than the bound (A.43) achieved by the argument in D. B. Madan (2010). Hence, the above line of arguments does not prove that the bid driver is balanced. In contrast, our proof relies on the dynamic translation invariance of the driver with which the desired lower bound can be achieved. Furthermore, in the constant liquidity setting this property enables us to arrive at a, compared to D. B. Madan (2010), simplified formula for the bid and ask price.

A.4 The induced nonlinear expectation

Due to the dynamic translation invariance of the bid and ask drivers, we can arrive at a backward recursive formula for the nonlinear expectations induced by the bid and ask drivers defined in the previous section, only depending on Q, Ψ and Γ. This is a new result that was unobtainable before, since the dynamic translation invariance of the drivers was not taken into account.

Proof of Theorem 1. By Theorem 3 and Lemma 3, there exist time-consistent and dynamically trans-

lation invariant nonlinear expectations $\mathcal{E}^{a,\Psi,\Gamma}$ and $\mathcal{E}^{b,\Psi,\Gamma}$ that are solutions to the backward recursion

$$\begin{aligned} \mathcal{E}(Q \,|\, \mathcal{F}_{t_k}) = Y_{t_k} &= \mathbb{E}[Y_{t_{k+1}} \,|\, \mathcal{F}_{t_k}] + F(\omega, t_k, Y_{t_k}, Z_{t_k}) \quad \forall t_k \in \mathcal{T}_0^{K-1} \\ Y_T &= Q, \end{aligned} \tag{A.44}$$

where F and $\mathcal{E}$ denote $F^{b,\Psi,\Gamma}$ and $\mathcal{E}^{b,\Psi,\Gamma}$ or $F^{a,\Psi,\Gamma}$ and $\mathcal{E}^{a,\Psi,\Gamma}$, respectively.

By Lemma 1 we have

$$Z_{t_k} M_{t_{k+1}} = Y_{t_{k+1}} - \mathbb{E}[Y_{t_{k+1}} \,|\, \mathcal{F}_{t_k}]. \tag{A.45}$$

Additionally, $F^{a,\Psi,\Gamma}$ and $F^{b,\Psi,\Gamma}$ are dynamically translation invariant. Therefore, for all $t_k \in \mathcal{T}_0^{K-1}$,

$$\begin{aligned} Y_{t_k} &= \mathbb{E}[Y_{t_{k+1}} \,|\, \mathcal{F}_{t_k}] + F^{b,\Psi,\Gamma}(\omega, t_k, Y_{t_k}, Z_{t_k}) \\ &= \mathbb{E}[Y_{t_{k+1}} \,|\, \mathcal{F}_{t_k}] + \mathbb{E}_{t_k}^{\psi^{\gamma_{t_k}}}[Y_{t_{k+1}} - \mathbb{E}[Y_{t_{k+1}} \,|\, \mathcal{F}_{t_k}]] \\ &= \mathbb{E}_{t_k}^{\psi^{\gamma_{t_k}}}[Y_{t_{k+1}}] \end{aligned} \tag{A.46}$$

and similarly for $F^{a,\Psi,\Gamma}$. By starting at $Y_T = Q$ and applying backward recursion we arrive at

$$\mathcal{E}^{b,\Psi,\Gamma}(Q \,|\, \mathcal{F}_{t_k}) = \mathbb{E}_{t_k}^{\psi^{\gamma_{t_k}}}[\mathbb{E}_{t_{k+1}}^{\psi^{\gamma_{t_{k+1}}}}[\cdots \mathbb{E}_{t_{K-1}}^{\psi^{\gamma_{t_{K-1}}}}[Q]]] = b_{t_k}^{\Psi,\Gamma}(Q) \tag{A.47}$$

and

$$\mathcal{E}^{a,\Psi,\Gamma}(Q \,|\, \mathcal{F}_{t_k}) = -\mathbb{E}_{t_k}^{\psi^{\gamma_{t_k}}}[\mathbb{E}_{t_{k+1}}^{\psi^{\gamma_{t_{k+1}}}}[\cdots \mathbb{E}_{t_{K-1}}^{\psi^{\gamma_{t_{K-1}}}}[-Q]]] = a_{t_k}^{\Psi,\Gamma}(Q), \tag{A.48}$$

which concludes the proof. □

B Binomial tree construction

Let $T > 0$ denote maturity. We consider the continuous time processes

$$dS_t = S_t(r - q)dt + S_t \sigma W_t^S \quad S_0 > 0, \tag{B.1}$$

$$d\gamma_t = \kappa(\theta - \gamma_t)dt + \nu\sqrt{\gamma_t} dW_t^\gamma, \quad \gamma_0 > 0, \tag{B.2}$$

where W^S and W^γ are correlated Brownian motions with $d\langle W^S, W^\gamma\rangle_t = \rho dt, \rho \in [-1,1]$. Here, σ denotes the implied volatility of a specific option's mid price. Using the transformation

$$x_t := \log S_t, \tag{B.3}$$

$$y_t := \frac{2}{\nu}\sqrt{\gamma_t} - \frac{\rho}{\sigma}x_t, \tag{B.4}$$

we get

$$dx_t = \mu_x dt + \sigma dW_t^S, \tag{B.5}$$

$$dy_t = \mu_y(x_t, y_t)dt + dW_t^\gamma - \varrho dW_t^S, \tag{B.6}$$

with

$$\mu_x := r - q - \frac{1}{2}\sigma^2, \tag{B.7}$$

$$\mu_y(x_t, y_t) := \left(\gamma_t(x_t, y_t)^{-\frac{1}{2}}\left(\frac{\kappa}{\nu}(\theta - \gamma_t(x_t, y_t))\right) - \frac{\nu}{4}\right) - \frac{\rho\mu_x}{\sigma}, \tag{B.8}$$

$$\gamma_t(x_t, y_t) := \left(\frac{\nu}{2}\left(y_t + \frac{\rho}{\sigma}x_t\right)\right)^2. \tag{B.9}$$

Defining the independent Brownian motions

$$B^x \equiv W^S, \quad B^y \equiv \frac{1}{\sqrt{1-\rho^2}}(W^\gamma - \varrho B^1), \tag{B.10}$$

we can rewrite the dynamics of x and y as

$$dx_t = \mu_x dt + \sigma dB_t^x, \tag{B.11}$$

$$dy_t = \mu_y(x_t, y_t)dt + \sqrt{1-\varrho^2}dB_t^y. \tag{B.12}$$

This decoupling allows us to have transition probabilities for x and y which only depend on whether x respectively y have previously moved up or down.

For a constant time step size $h := \frac{T}{K}$, where T is the maturity and K the number of time steps,

the processes x and y are discretized, for $k = 0, ..., K$, as random walks:

$$X_k^{(K)} := x_0 + \sqrt{h}\sigma \sum_{i=1}^{k} \xi_i^X, \tag{B.13}$$

$$Y_k^{(K)} := y_0 + \sqrt{h(1-\varrho^2)} \sum_{i=1}^{k} \xi_i^Y, \tag{B.14}$$

where $(\xi_i^X)_{i=1}^K$ and $(\xi_i^Y)_{i=1}^K$ are independent random variables with values in $\{\pm 1\}$. We use the convention that $X_0^{(K)} \equiv x_0$ and $Y_0^{(K)} \equiv y_0$. In addition, we define the filtration

$$\mathcal{F}_k := \sigma(\xi_1^X, \xi_1^Y, ..., \xi_k^X, \xi_k^Y), \tag{B.15}$$

for $k = 1, ..., K$ and $\mathcal{F}_0 = \{\emptyset, \Omega\}$, where Ω is the obvious state space. The probabilities $p_k := \mathbb{P}_{k-1}(\xi_k^X = 1)$ and $q_k := \mathbb{P}_{k-1}(\xi_k^Y = 1)$ are determined via moment matching, which guarantees the weak convergence of $(X^{(n)}, Y^{(n)})$ to (x, y) (see, e.g., Ethier and Kurtz (2009) and Akyıldırım et al. (2014)). In particular, the first and second moments must satisfy

(i) $\mathbb{E}[X_k^{(K)} - X_{k-1}^{(K)} \,|\, \mathcal{F}_{k-1}] = \mu_x h + o(h)$,

(ii) $\mathbb{E}[Y_k^{(K)} - Y_{k-1}^{(K)} \,|\, \mathcal{F}_{k-1}] = \mu_y(X_{k-1}^{(K)}, Y_{k-1}^{(K)})h + o(h)$,

(iii) $\mathbb{E}[(X_k^{(K)} - X_{k-1}^{(K)})^2 \,|\, \mathcal{F}_{k-1}] = \sigma^2 h + o(h)$,

(iv) $\mathbb{E}[(Y_k^{(K)} - Y_{k-1}^{(K)})^2 \,|\, \mathcal{F}_{k-1}] = (1-\varrho^2)h + o(h)$.

The second moment conditions are fulfilled by construction. The first moment conditions lead to:

$$p_k = \mathbb{P}_{k-1}(\xi_k^X = 1) = \frac{1}{2}\left(1 + \frac{\mu_x\sqrt{h}}{\sigma}\right) \tag{B.16}$$

$$q_k = \mathbb{P}_{k-1}(\xi_k^Y = 1) = \frac{1}{2}\left(1 + \frac{\mu_y(X_{k-1}^{(K)}, Y_{k-1}^{(K)})\sqrt{h}}{\sqrt{1-\varrho^2}}\right). \tag{B.17}$$

As in, e.g., Akyıldırım et al. (2014), we need to truncate the transition probabilities such that they

take values in the unit interval. Therefore, we use the modified definitions:

$$p_k = \max\left\{0, \min\left\{1, \frac{1}{2}\left(1 + \frac{\mu_x\sqrt{h}}{\sigma}\right)\right\}\right\}, \tag{B.18}$$

$$q_k = \max\left\{0, \min\left\{1, \frac{1}{2}\left(1 + \frac{\mu_y(X_{k-1}^{(K)}, Y_{k-1}^{(K)})\sqrt{h}}{\sqrt{1-\varrho^2}}\right)\right\}\right\}. \tag{B.19}$$

Optimal Conic Execution Strategies with Stochastic Liquidity

Markus Leippold and Steven Schärer

Abstract

We develop the Conic Finance framework for optimal execution of a large portfolio in an illiquid market. We first extend the classical optimal execution results by considering nonlinear temporary market impact functions motivated by the theory of Conic Finance. The second extension is to consider a stochastic exogenous liquidity component to capture the base illiquidity of a market. We furthermore consider stochastic volatility and endogenous liquidity effects. We derive the Bellman equations for the risk-neutral and exponential objective functions and analyze various aspects of our model using a stylized example.

JEL Classification: C61, G11, G12

Keywords: Optimal trade execution, Conic Finance, stochastic liquidity, dynamic programming

1 Introduction

Optimal execution theory is concerned with a problem faced in practice by many large portfolio managers and other institutional investors. Namely, executing a large buy or sell order of an asset can adversely impact its market price. As a result, traders may want to split their orders into smaller parts to be executed over several days (or nowadays, minutes or less) in the hope of reducing the price impact. This price impact is alleviated at the cost of price uncertainty over the scheduled execution period.

A solution to the trade-off of implementing an execution schedule for a large trade to reduce its market impact while being exposed to significant asset movements independent of the trading activity was first proposed in the seminal papers of Bertsimas and Lo (1998) and Almgren and Chriss (1999) and Almgren and Chriss (2001). In the model of Almgren and Chriss, the asset price follows an arithmetic Brownian motion that is impacted by both permanent and temporary liquidity effects.[28] Both liquidity components are caused by the trading activity of the investor buying or selling a large position. The temporary effects are assumed to prevail only for as long as it takes to execute a single trade. In contrast, the permanent effects accumulate over the whole trading period. Optimal trading strategies are derived in both papers by minimizing the implementation shortfall, i.e., the difference between the market price observed at the time the decision is made to buy or sell and the average price resulting from the trading process. But while Bertsimas and Lo (1998) only minimize the expected implementation shortfall, Almgren and Chriss (2001) also take into account its variance.[29]

In most of the literature, the liquidity effects described above are assumed to be linear. However, as Almgren (2003) argues, this assumption is not very realistic. In particular, the more of an asset needs to be traded over a shorter amount of time, the more disproportionate the effect on the resulting execution price. Therefore, we suggest to model the nonlinear temporary market impact within a

[28]The choice of an arithmetic Brownian motion may lead to negative prices. However, for realistic parameter values, the probability for negative prices becomes negligible.

[29]The model of Bertsimas and Lo (1998) assumes a geometric Brownian motion for the underlying stock price process, overcoming some of the deficiencies of the Almgren and Chriss (1999) and Almgren and Chriss (2001) model. However, this assumption comes at the cost of increased computational complexity for the derivation of optimal execution strategies.

Conic Finance framework, which provides a way of computing bid and ask prices depending on the current liquidity of a market.[30]

Conic Finance was first introduced by D. B. Madan and Cherny (2010a). It postulates the existence of a central counterparty which executes all trades that are acceptable to it, where the concept of acceptability is closely linked to coherent risk measures and in particular distortion risk measures. This framework gives rise to bid and ask prices of assets, which can be thought of as distorted expectations.[31] Since bid-ask spreads are crucial for optimal execution strategies, Conic Finance lends itself well to our problem.

In Figure 1 we plot the intraday relative bid-ask spreads for the IBM stock traded on the NYSE, averaged over the month of January 2017. This spread is often assumed to be constant in the optimal execution literature as, e.g., in Almgren and Chriss (2001) and Cheridito and Sepin (2014). Such an assumption, however, might be too restrictive. First of all, we observe that the median of bid-ask spreads follows a pattern that corresponds to what empirical studies call the "J-shape", i.e., high spreads shortly after market opening and relatively constant spreads until market closing. The J-shape is a phenomenon common to many intraday markets and for different variables, such as volatility and volume.[32] Second, the 90% confidence band indicates that the spreads might vary substantially, even for a liquid stock like IBM. Lastly, when we look at a single day, we observe that bid-ask spreads might even spike over a trading day. Hence, we take this observation as a motivation to enrich our model with a separate stochastic parameter, allowing us to capture this base illiquidity of the market.

The models of Bertsimas and Lo (1998) and Almgren and Chriss (2001) have been extended in

[30] We will keep in line with the current literature in that we model the permanent market impact as a linear function of the trading volume. Indeed, as has been discovered by Huberman and Stanzl (2004) and Gatheral (2010), models with permanent market impact such as Almgren and Chriss (2001) allow for dynamic arbitrage if the market impact function is not linear. For example, there exists a strategy of buy and sell orders such that the final position in the asset equals the initial position, but the cash account at the end of the trading period is expected to be greater or equal than the initial one. This has since been made less strict by Guéant (2013), whose approach could also be added to our model.

[31] The fit of these bid and ask prices on real data has been studied empirically, for example for daily close prices of S&P 500 vanilla options in Albrecher et al. (2013). The model proposed by D. B. Madan and Cherny (2010a) has been applied to various use cases and was extended in multiple ways such as the extension to a discrete-time model in D. B. Madan (2010) and to stochastic liquidity in the same framework by Leippold and Schärer (2017).

[32] See, e.g., Zwergel and Heiden (2014) and the references contained therein for a literature overview. Other typical shapes include the "U-shape", for which the spreads would again increase shortly before market closing, or the "W"-shape with spikes at the opening, closing, and around midday.

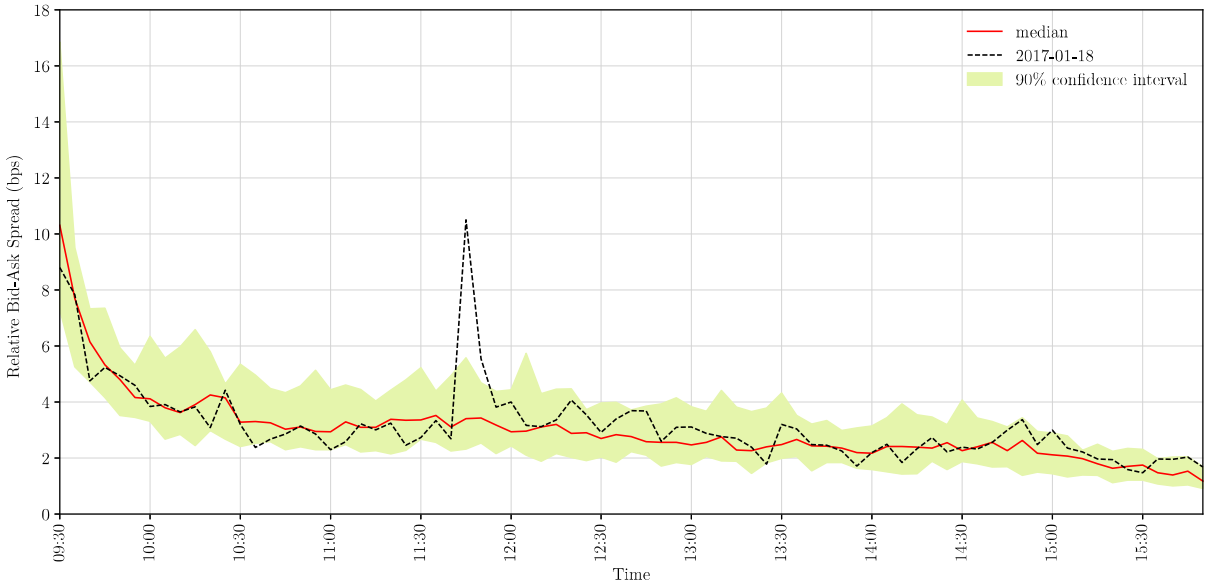


Figure 1. We plot the relative bid-ask spreads for the IBM stock traded on the NYSE during January 2017. For each trading day in January, we split the day into five minute intervals from 09:30 to 16:30 local time. In each time interval, we calculate the average relative bid-ask spread. We then calculate the median (red line) and 90% confidence interval (green area) over the whole time interval. We also display the relative bid-ask spreads for a single day, 18.01.2017 (dashed black line). The relative bid-ask spread for each tick t is calculated as $s_t = 2(b_t - a_t)/(a_t + b_t)$, where a_t and b_t are the ask and bid price for tick t, respectively. The data is from the TAQ database accessed via WRDS.

many directions and we refer to Gatheral and Schied (2013) for an overview. All of these papers (including ours) try to minimize some measure of the implementation shortfall using market orders, assuming a permanent and temporary price impact component in a discrete-time model. Each of these features and assumptions have been extended in some way or another in the growing body of literature on optimal execution. An alternative framework, derived by modeling the limit order book was first proposed by Obizhaeva and J. Wang (2012) and followed up by research from Alfonsi, Fruth, et al. (2010), which also considers non-linear price impact functions. Curato et al. (2016) also consider non-linear market impact though in a market order framework. These models also generalize the market impact of trading to not be permanent, but transient, as do Gatheral, Schied, and Slynko (2012) and Curato et al. (2016). Optimal execution in a multi-asset context is considered, e.g., in Schied et al. (2010) and Lin and Fahim (2017). We refer to Guéant (2016) for both a broader literature overview and various extensions such as target functions different from implementation shortfall, multi-asset portfolios and limit orders, where we especially refer to Guéant and Lehalle (2015) for the last topic. Price impact models can also be used to study the implications of the co-existence of traditional exchanges and dark pools as in, e.g., P. Kratz and Schöneborn (2018) and the problem of hedging in illiquid markets by Guéant and Pu (2015). Fukasawa and Stadje (2018) consider a similar market setup in which the price for a certain amount of stocks is provided by g-expectations and derive perfect dynamic hedging strategies.

The rest of the paper is structured as follows. In Section 2, we introduce the Conic Finance model of D. B. Madan and Cherny (2010a) and extend it to a filtered framework with random liquidity, closely resembling the setup of Leippold and Schärer (2017) though with only one time step. Section 3 introduces the probability space and asset price model with price impact considerations. The optimal execution problem is defined in Section 4 and solved for two different utility measures. In Section 5, we present a toy example based on the parameters of Cheridito and Sepin (2014) and further analyze our model. Section 6 concludes. All proofs are delegated to the appendix.

2 Option pricing with random liquidity

In this section, we introduce the main concepts of Conic Finance that we need so that we can adapt the theory to formulate the optimal execution problem.

2.1 The model of D. B. Madan and Cherny (2010a)

D. B. Madan and Cherny (2010a) proposed a model for bid and ask prices of instruments with terminal cash flows in a single time-step framework. It is based on the idea of *acceptability*, i.e., a postulated central counterparty executes all acceptable trades. This central counterparty can be thought of as the market. Then, the bid price of an instrument is the highest price at which the market buys from the seller such that the resulting net trade is still acceptable to the market, and similar for the ask price.

To clarify this concept, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. By $\mathbb{P}$ we denote the reference probability measure, which is assumed to be risk-neutral. In complete and perfectly liquid markets, $\mathbb{P}$ is unique and cash flows $Q \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ are considered acceptable if their risk-neutral expectation are non-negative, i.e., $\mathbb{E}^{\mathbb{P}}[Q] \geq 0$. As a result, the bid and ask price coincide and are equal to the risk-neutral expectation.[33] In our incomplete and illiquid market, there generally is a non-zero spread between bid and ask prices, as is observed in reality.

In the model of D. B. Madan and Cherny (2010a), the central counterparty evaluates potential trades not by merely calculating the expected value of net cash flows with respect to $\mathbb{P}$, but with respect to a distorted measure. If the counterparty is supposed to buy a cash flow, for example, it overweights the left side of the distribution (the probability of losses) and underweights the right side (the probability of profits). Formally, the distorted expectations are defined as follows.

Definition 16 (Distorted expectation)**.** *Let ψ be a concave distortion function*[34] *and $Q \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$*

[33]Since the bid price b of a discounted cash flow Q is the largest price $x \in \mathbb{R}$ that the central counter party is willing to pay such that the net cash flow from the point view of the central counter party, $Q - x$, is acceptable, i.e., $\mathbb{E}^{\mathbb{P}}[Q - x] \geq 0$. This of course implies that $b = \mathbb{E}^{\mathbb{P}}[Q]$. The same argument for the ask price a, but with a net cash flow of $x - Q$ results in $a = \mathbb{E}^{\mathbb{P}}[Q] = b$.

[34]A function $\psi : [0, 1] \longrightarrow [0, 1]$ is a distortion function if and only if it is monotone, $\psi(0) = 0$, and $\psi(1) = 1$.

a future discounted cash flow. The distorted expectation of Q is denoted by $\mathbb{E}^{\psi}[Q]$ and defined as

$$\mathbb{E}^{\psi}[Q] := \int_0^{\infty} (1 - \psi(\mathbb{P}(Q \leq x)))dx - \int_{-\infty}^{0} \psi(\mathbb{P}(Q \leq x))dx \tag{2.1}$$

Generally, $\psi \circ \mathbb{P}$ is no longer a probability measure. Nevertheless, the theory of Choquet integrals[35] can be used to define distorted expectations, which intuitively corresponds to the idea of over- and underweighting different areas of the probability distribution as described above. Distorting the probability measure implies that distorted expectations are no proper expectations anymore. In particular, they are non-linear, i.e., $\mathbb{E}^{\psi}[Q_1 + Q_2] \neq \mathbb{E}^{\psi}[Q_1] + \mathbb{E}^{\psi}[Q_2]$ for $Q_1, Q_2 \in L^{\infty}(\Omega, \mathcal{F}, \mathbb{P})$ in general. However, distorted expectations still share some properties such as monotonicity, positive homogeneity, and translation invariance with the usual expectation operator $\mathbb{E}^{\mathbb{P}}[\cdot]$.

The central counterparty is postulated to apply a certain distortion function to the reference probability measure $\mathbb{P}$, depending on how *liquid* it is. The less liquid the market, the bigger the distortion with respect to $\mathbb{P}$. For that, consider a family $\Psi = (\psi^z)_{z \geq 0}$ of concave distortion functions. This family is pointwise increasing in z, i.e., $\psi^{z_1}(\cdot) \leq \psi^{z_2}(\cdot)$ if and only if $z_1 \leq z_2$. We also assume that $\psi^0 \equiv id$ the identity function. The liquidity level of the market is then measured by a non-negative constant variable γ according to which the market selects the distortion function ψ^{γ} among Ψ which it then applies to $\mathbb{P}$. Thereby, $\gamma = 0$ corresponds to a perfectly liquid market as $\psi^0 \circ \mathbb{P} = \mathbb{P}$ and the larger γ, the less liquid the market is considered to be.

A discounted net cash flow Q is then considered acceptable to the central counterparty with a family of distortion functions Ψ and liquidity level γ, if and only if $\mathbb{E}^{\psi^{\gamma}}[Q] \geq 0$. Since the market is assumed to be competitive, we have the following expressions for bid and ask prices in this single-period model.

Definition 17 (One-period bid and ask prices)**.** *Let $\Psi = (\psi^z)_{z \geq 0}$ be a family of pointwise increasing concave distortion functions and $\gamma \geq 0$. The bid price of a cash flow $Q \in L^{\infty}(\Omega, \mathcal{F}, \mathbb{P})$ is given by $\mathbb{E}^{\psi^{\gamma}}[Q]$ and the ask price is $-\mathbb{E}^{\psi^{\gamma}}[-Q]$.*

[35] See Choquet (1953)

Due to the properties of the Choquet integral, it holds that

$$\mathbb{E}^{\psi^\gamma}[Q] \leq \mathbb{E}^{\mathbb{P}}[Q] \leq -\mathbb{E}^{\psi^\gamma}[-Q], \quad \forall Q \in L^\infty, \tag{2.2}$$

and hence there is a non-negative bid-ask spread. We end this section by introducing two examples of concave distortion functions that we need for our example that we present in Section 5. Other examples of distortion functions include the MinMaxVar distortion used in Cherny and D. B. Madan (2009) and the EssSupExp distortion used in Bannör and Scherer (2014).

Example 1 (Wang distortion)**.** *The Wang distortion*[36] *is given by*

$$\psi^z_{\text{Wang}}(u) := \Phi(\Phi^{-1}(u) + z), \quad \forall u \in [0,1], z \geq 0, \tag{2.3}$$

where Φ *is the cumulative distribution function of the standard normal distribution and* Φ^{-1} *the inverse thereof.*

Example 2 (CVaR distortion)**.** *The Conditional Value-at-Risk (CVaR)*[37] *at level* α *can be expressed as a distorted expectation,*

$$CVaR_\alpha(Q) = -\mathbb{E}^{\psi^z_{\text{CVaR}}}[Q], \quad \forall Q \in L^\infty, \tag{2.4}$$

for the concave distortion function

$$\psi^z_{\text{CVaR}}(u) := \min\left\{\frac{u}{1-\varphi(z)}, 1\right\}, \quad \forall u \in [0,1], \tag{2.5}$$

and $\alpha = 1 - \varphi(z)$*, for a sigmoid function* φ*, e.g.,*

$$\varphi(x) := \frac{x}{\sqrt{1+x^2}} \quad \forall x \in \mathbb{R}_+. \tag{2.6}$$

For our framework, we will use φ *to transform the confidence level* $\alpha \in (0,1]$ *to a liquidity level* $z \in [0, \infty)$*.*

[36] See S. S. Wang (2000).

[37] See, e.g., Föllmer and Schied (2011).

2.2 Distorted conditional expectations and random liquidity

For our optimal execution framework, we need to extend the previous model of D. B. Madan and Cherny (2010a). In particular, when placing market orders, the resulting bid or ask prices are only known after execution has occurred. Therefore, they need to be modeled as random variables. We propose to model this by letting the market liquidity parameter γ become a random variable. And since we will be working in a multi-period model, we will also introduce the concept of distorted conditional expectations. Both extensions are based on the theory developed for multi-period stochastic liquidity model by Leippold and Schärer (2017). First, we introduce the concept of a state-dependent distortion function.

Definition 18 (Concave state-dependent distortion function)**.** *Let $\mathcal{H} \subset \mathcal{F}$ a sub-σ-algebra. A function $\psi : \Omega \times [0,1] \longrightarrow [0,1]$ is called a $\mathcal{H}$-measurable concave state-dependent distortion function if and only if for all $\omega \in \Omega$, $\psi(\omega, \cdot)$ is a concave distortion function and for all $u \in [0,1]$, $\psi(\cdot, u)$ is $\mathcal{H}$-measurable.*

The above definition allows us to introduce the equivalent of the distorted expectation in Definition 16, the distorted conditional expectation for which we will denote by $\mathbb{P}(\cdot \,|\, \mathcal{H})$ the conditional probability with respect to a sub-σ-algebra $\mathcal{H} \subset \mathcal{F}$[38]

Definition 19 (Distorted conditional expectation)**.** *Let $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ sub-σ-algebras of $\mathcal{F}$ and ψ be a $\mathcal{G}$-measurable concave state-dependent distortion function. The function $\mathbb{E}^{\psi}[\cdot \,|\, \mathcal{H}] : L^{\infty}(\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow L^{\infty}(\Omega, \sigma(\mathcal{H} \cup \mathcal{G}), \mathbb{P})$, defined $\forall Q \in L^{\infty}(\Omega, \mathcal{F}, \mathbb{P})$ and $\forall \omega \in \Omega$ as*

$$\mathbb{E}^{\psi}[Q \,|\, \mathcal{H}](\omega) := -\int_{-\infty}^{0} \psi(\omega, \mathbb{P}(Q \leq x \,|\, \mathcal{H})(\omega))dx + \int_{0}^{\infty} 1 - \psi(\omega, \mathbb{P}(Q \leq x \,|\, \mathcal{H})(\omega))dx, \tag{2.7}$$

is called distorted conditional expectation.

Note that for each $\omega \in \Omega$, $\mathbb{E}^{\psi}[\cdot \,|\, \mathcal{H}](\omega)$ is a distorted expectation as defined in the previous section. The distorted conditional expectations are also

(i) monotone, i.e., $\mathbb{E}^{\psi}[Q^1 \,|\, \mathcal{H}] \leq \mathbb{E}^{\psi}[Q^2 \,|\, \mathcal{H}]$, for $Q^1 \leq Q^2$ $\mathbb{P}$-a.s.,

[38] We assume enough regularity on the filtered probability space that a regular conditional distribution can be constructed.

(ii) adapted, i.e., $\mathbb{E}^{\psi}[Q \,|\, \mathcal{H}] = Q$, if Q is $\mathcal{H}$-measurable,

(iii) dynamic translation invariant, i.e., $\mathbb{E}^{\psi}[Q + q \,|\, \mathcal{H}] = \mathbb{E}^{\psi}[Q \,|\, \mathcal{H}] + q$, for Q $\mathcal{F}$-measurable and q $\mathcal{H}$-measurable.

as is easily verified. Contrary to the usual conditional expectation, $\mathbb{E}^{\psi}[\cdot \,|\, \mathcal{H}](\omega)$ is only $\sigma(\mathcal{H} \cup \mathcal{G})$-measurable. We will use the concept of distorted conditional expectations to define the bid and ask prices in our optimal execution framework. to this end, we simplify notation and write, for a random variable γ,

$$\psi^{\gamma}(\omega, u) := \psi^{\gamma(\omega)}(u) \quad \forall \omega \in \Omega, \forall u \in [0, 1]. \tag{2.8}$$

In particular it holds that ψ^{γ} is $\mathcal{H}$-measurable if and only if γ is $\mathcal{H}$-measurable.

3 The asset price and market impact model

Let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n=0}^N, \mathbb{P})$ be a finite filtered probability space satisfying the usual conditions. The filtration will be constructed explicitly at the end of the section.

3.1 Stock process and execution strategy

Our starting point is the liquidation of a position $X > 0$ in a certain stock.[39] Let $T > 0$ be the liquidation horizon, usually in the minutes to hours range. Since we consider a discrete-time model, we partition the time from now to the end of the liquidation period into $N > 0$ time intervals with constant length $\Delta t := T/N$. In particular, we have time steps $0 = t_0 < \cdots < t_n = n\Delta t < \cdots < t_N = T$, for $n \in \{1, ..., N-1\}$.

Let $(x_n)_{n=0}^N$ denote the execution strategy, i.e., the number of remaining stocks at each time step t_n. By the definition of the problem, we require the execution strategy to start at the initial position, $x_0 = X$, there being no position left at the end of the liquidation window, $x_N = 0$, and the

[39] We could also consider the problem of building up a position in a certain stock. We will mention the necessary changes to the model where appropriate.

position to only decrease over time, $x_{n-1} \geq x_n$, for all $n = 1, ..., N$.[40] It is sometimes easier to work with $y_n := x_{n-1} - x_n$, the number of stocks sold in period $(t_{n-1}, t_n]$. By construction it holds that $\sum_{n=1}^{N} y_n = X$.

The stock price process $S = (S_n)_{n=0}^{N}$ is assumed to follow[41]

$$S_n = S_{n-1} + \sigma_n \xi_n - c y_n, \quad \forall n = 1, ..., N, \tag{3.1}$$

for an initial price $S_0 > 0$.[42] We hereby assume that there is a linear permanent price impact of selling stocks, measured by $c > 0$. The stock process is also driven by a stochastic volatility process $(\sigma_n)_{n=0}^{N}$ and process $(\xi_n)_{n=0}^{N}$ with independently and identically distributed (iid) innovations, with each innovation being drawn from a normal distribution with mean zero and volatility $\sqrt{\Delta t}$, i.e., $\xi_n \sim \mathcal{N}(0, \Delta t)$. We denote the stock price before the permanent price impact, or equivalently, the stock price when the position has not changed, by $\widetilde{S}_n := S_{n-1} + \sigma_n \xi_n$, for all $n = 1, ..., N$. In particular, $\mathbb{P}$ is risk-neutral with respect to $\widetilde{S}_n$.

3.2 Definition of the bid price

Stocks are sold at the market's bid price. The theory of Conic Finance as introduced by D. B. Madan and Cherny (2010a) allows to calculate the bid and ask prices of cash flows taking into account the market liquidity. In this framework, however, we are interested in the bid price of a stock which does not have a terminal cash flow. Due to the short time horizon and the assumption of zero interest rate and dividend yield, we can equate selling one unit of stock at time t_n, with selling an artificial forward contract with maturity t_{n+1} and delivery price 0 (and hence payoff $\widetilde{S}_{n+1}$).[43]

[40] We explicitly do not allow for strategies that include buying periods to stay consistent with Almgren (2003) and Cheridito and Sepin (2014) and to improve the calculation time of the optimal execution strategies. We analyze the implications of removing this restriction in Appendix B and find, at least for the examples considered in this paper, no difference in optimal execution strategies and in particular no evidence of transaction-triggered price manipulation as defined by Alfonsi, Schied, et al. (2012).

[41] In particular, the risk-free rate and dividend yields are assumed to be zero.

[42] The stock price has to remain positive at all times. Due to the short time frame considered this should not be an issue even with a process that can become negative given enough time.

[43] We could also buy a short forward with the same maturity and delivery price. The two approaches are equivalent because, by the definition of conditional distorted non-linear expectations, $b_{n+1}^{\Psi,\widetilde{\Gamma}} = \mathbb{E}_n^{\psi^{\widetilde{\gamma}_{n+1}}}[\widetilde{S}_{n+1}] = -\mathbb{E}_n^{\psi^{\widetilde{\gamma}_{n+1}}}[-\widetilde{S}_{n+1}] = -a_{n+1}^{\Psi,\widetilde{\Gamma}}$.

There are at least two different types of liquidity effects. The first is the exogenous market liquidity, which is independent of any single trader's activity. We model this effect using an adapted stochastic process $\gamma = (\gamma_n)_{n=0}^N$. This liquidity, which is often modeled as a constant (or ignored altogether) in the classical optimal execution models,[44] captures the "base" liquidity of the market, which is observable at the time of placing an order by, e.g., looking at the prevalent bid-ask spreads. The second effect is due to the placement of a market order, which in the case of a large order often results in a lower (average) bid price, as the available liquidity in the order book is exhausted. In particular, following the approach of Cheridito and Sepin (2014), we define the endogenous liquidity effect of an order of the size y_{n+1} at time t_n by $\eta_{n+1}y_{n+1}$. At the time of placing an order, it is therefore not known how much the endogenous liquidity process $\eta = (\eta_n)_{n=0}^N$ will influence the resulting price.

Different than in the framework of D. B. Madan and Cherny (2010a), the bid price therefore not only depends on an adapted exogenous market liquidity process γ, but also on the future realization of the endogenous liquidity process η, which is not known at the time a market order is placed. We therefore define the bid price, now a random variable, using distorted conditional expectations stated in Definition 8, where for ease of notation we denote

$$\mathbb{E}_n^{\psi}[\cdot] := \mathbb{E}^{\psi}[\cdot \,|\, \mathcal{F}_n]. \tag{3.2}$$

Definition 20 (Trade-size dependent bid price)**.** *The bid price per stock received for selling y_{n+1} stocks in the time interval $(t_n, t_{n+1}]$ is*

$$b_{n+1}^{\Psi,\gamma,\eta} := \mathbb{E}_n^{\psi^{\gamma_n+\eta_{n+1}y_{n+1}}}[\widetilde{S}_{n+1}]. \tag{3.3}$$

We note that the bid price $b_{n+1}^{\Psi,\gamma,\eta}$ is $\mathcal{F}_{n+1}$-measurable, and in particular $\sigma(\mathcal{F}_n, \eta_{n+1})$-measurable. Hence, by distorting the expectation by an appropriate function driven by two components, an exogenous market liquidity process γ and a market impact process η, we can express the bid price in a convenient way to formulate the optimal execution problem in an illiquid market.

Remark 3. *The currently observed bid price at time t_n, not influenced by any potential market*

[44] See, e.g., parameter ϵ in Almgren and Chriss (2001).

orders, is given by

$$\mathbb{E}_n^{\psi^{\gamma_n}}[\widetilde{S}_{n+1}], \tag{3.4}$$

which of course is $\mathcal{F}_n$-measurable. The observed bid prices can be used to calibrate the processes σ and γ.

Remark 4. *If we are interested in building up a stock position from zero to X, it is the ask price per stock that is relevant. Following the same lines of arguments, the ask price is given by*

$$a_{n+1}^{\Psi,\gamma,\eta} := -\mathbb{E}_n^{\psi^{\gamma_n+\eta_{n+1}|y_{n+1}|}}[-\widetilde{S}_{n+1}]. \tag{3.5}$$

Similarly as in, e.g., Almgren and Chriss (2001) we need to consider the absolute order size $|y_{n+1}|$ when calculating the temporary market impact.

In the setup outlined above, the filtration $(\mathcal{F}_n)_{n=0}^N$ is given by the σ-algebras spanned by S_n, σ_n, η_n, and γ_n. We assume ξ_n is independent of $\sigma(\mathcal{F}_{n-1}, \sigma_n)$. Furthermore, for the remainder of the paper, we assume that $(\sigma_n, \eta_n, \gamma_n)$ is a Markov chain with finite state space $V \subseteq \mathbb{R}_+^3$ and transition probabilities

$$p_{n-1}^{vw} := \mathbb{P}\left[(\sigma_n, \eta_n, \gamma_n) = w \,|\, (\sigma_{n-1}, \eta_{n-1}, \gamma_{n-1}) = v\right], \quad \forall v, w \in V. \tag{3.6}$$

We will denote the elements of a realization v at any time step $n = 0, ..., N$ of the Markov chain as $v = (v_\sigma, v_\eta, v_\gamma)$ and by

$$\mathbb{E}_{n,v}[\cdot] := \mathbb{E}^{\mathbb{P}}[\cdot \,|\, \mathcal{F}_n, (\sigma_n, \eta_n, \gamma_n) = v] \tag{3.7}$$

the conditional expectations given the realization v of the Markov chain $(\sigma_n, \eta_n, \gamma_n)_{n=0}^N$ at time n and by

$$\mathbb{E}_{n,v}^{\psi}[\cdot] := \mathbb{E}^{\psi}[\cdot \,|\, \mathcal{F}_n, (\sigma_n, \eta_n, \gamma_n) = v] \tag{3.8}$$

the distorted expectation equivalent.

4 Determining optimal execution strategies

The goal behind optimal execution is to find a strategy $x = (x_n)_{n=0}^N$ that minimizes the implementation shortfall of selling a position X in the stock. This measure, also called implementation cost

or slippage, is the difference between the book value of the position and the accumulated bid prices received over the course of the trading activity. Hence, in our Conic Finance framework, we can define the implementation shortfall of selling a position X simply as

$$C(x) := XS_0 - \sum_{n=1}^{N}(x_{n-1} - x_n)b_n^{\Psi,\gamma,\eta}. \tag{4.1}$$

We note that, consistent with previous literature, we define $C(x)$ based on the unobservable S_0 for simplicity. Alternatively and without changing the resulting optimal execution strategies, we could have replaced S_0 by, e.g., the best bid price at time zero which would actually be observable on the market, or the average between the bid and ask price. Since $C(x)$ depends on the future realizations of σ, η, γ, and ξ, the investor needs to choose a measure with respect to which the implementation shortfall is minimized. Note that, at each point in time t_n, $n = 0, ..., N-1$, it is only decided how much to sell in the next trading period. After that, the current market conditions are observed again, before deciding on the next period's trading volume.

By the dynamic translation invariance property of distortion risk measures and some algebra[45], we can write the implementation shortfall in equation (4.1) given the stock price dynamics in equation (3.1) as

$$C(x) = \frac{c}{2}X^2 - \sum_{n=1}^{N}(x_{n-1} - x_n)^2\frac{c}{2} + (x_{n-1} - x_n)\mathbb{E}_{n-1}^{\psi^{\gamma_{n-1}+\eta_n(x_{n-1}-x_n)}}[\sigma_n\xi_n] + x_n\sigma_n\xi_n. \tag{4.2}$$

Below, we give three examples in which we can derive the implementation shortfall in closed form.

Example 3 (Constant volatility and Wang distortion)**.** *One particular example for which the distorted nonlinear expectation can be solved in closed-form is the constant volatility $\sigma_n \equiv \sigma$ and Wang distortion case given in Example 1, $\Psi = \Psi^{Wang}$. By definition of the Wang distortion function it holds that*

$$\mathbb{E}_{n-1}^{\psi_{Wang}^z}[\xi] = -\sqrt{\Delta t}z \tag{4.3}$$

[45] We refer to the appendix for the proof.

for a random variable $\xi \sim \mathcal{N}(0, \Delta t)$ *and* $z \geq 0$*. Hence,*

$$\mathbb{E}_{n-1}^{\psi^{\gamma_{n-1}+\eta_n y_n}}[\sigma\xi_n] = -\sigma\sqrt{\Delta t}(\gamma_{n-1} + \eta_n(x_{n-1} - x_n)) \tag{4.4}$$

and thus the implementation shortfall simplifies to

$$C(x) = \frac{c}{2}X^2 + \sum_{n=1}^{N}(x_{n-1} - x_n)^2(\sigma\eta_n\sqrt{\Delta t} - \frac{c}{2}) + (x_{n-1} - x_n)\gamma_{n-1}\sigma\sqrt{\Delta t} - x_n\sigma\xi_n. \tag{4.5}$$

This also shows that as long as $\gamma_n \equiv \gamma \geq 0$ *is constant, the optimal trading strategies are independent of its value as in that case the implementation shortfall can be written as*

$$C(x) = \frac{c}{2}X^2 + \gamma\sigma\sqrt{\Delta t}X + \sum_{n=1}^{N}(x_{n-1} - x_n)^2(\sigma\eta_n\sqrt{\Delta t} - \frac{c}{2}) - x_n\sigma\xi_n. \tag{4.6}$$

Example 4 (Cheridito and Sepin (2014) with constant volatility)**.** *If additionally to the assumptions in Example 3 (*$\sigma_n \equiv \sigma$ *and* $\Psi = \Psi^{Wang}$*), we assume that* $\gamma_n \equiv 0$*, we arrive at the same implementation shortfall as Cheridito and Sepin (2014) under constant volatility,*

$$C(x) = \frac{c}{2}X^2 + \sum_{n=1}^{N}(x_{n-1} - x_n)^2(\widetilde{\eta}_n - \frac{c}{2}) - x_n\sigma\xi_n \tag{4.7}$$

where $\widetilde{\eta}_n = \sigma\eta_n\sqrt{\Delta t}$ *is the temporary stochastic liquidity process as in Cheridito and Sepin (2014). In our framework, this process also depends on the volatility of the underlying since our bid price is a nonlinear distorted expectation. As such, our model is a generalization of the constant volatility version of Cheridito and Sepin (2014). Therefore, we are also able to recover the classical results of Bertsimas and Lo (1998) and Almgren and Chriss (2001).*

Example 5 (Almgren (2003))**.** *Almgren (2003) proposes to make the temporary market impact nonlinear by defining, in the notation of our paper,*

$$b_{n-1}^{\Psi,\gamma,\eta} = S_{n-1} - \widehat{\eta}y_n^k, \tag{4.8}$$

for some temporary market liquidity parameter $\widehat{\eta}$ *and an additional parameter* $k > 0$ *controlling the*

nonlinearity of the temporary market impact. This specification can be seen to be a special case of our model by assuming $\gamma_n \equiv 0$, $\sigma_n \equiv \sigma > 0$ *and* $\eta_n \equiv \eta > 0$ *and introducing a new concave distortion function,*

$$\psi_k^z(u) := \psi_{\text{Wang}}^{z^k}(u) = \Phi(\Phi^{-1}(u) + z^k), \quad \forall u \in [0,1], z \geq 0. \tag{4.9}$$

The family $\Psi_k := (\psi_k^z)_{z \geq 0}$ *is pointwise increasing and we find that* $b_n^{\Psi_k,\gamma,\eta} = S_{n-1} - \widehat{\eta} y_n^k$ *with*

$$\widehat{\eta} = \sigma \eta^{1/k} \sqrt{\Delta t}. \tag{4.10}$$

Then, the implementation shortfall is given by

$$C(x) = \frac{c}{2} X^2 - \sum_{n=1}^{N} (x_{n-1} - x_n)^2 \frac{c}{2} - (x_{n-1} - x_n)^{k+1} \widehat{\eta} + x_n \sigma_n \xi_n. \tag{4.11}$$

4.1 Risk-neutral objective

In the classical work of Bertsimas and Lo (1998), the authors rely on a risk-neutral objective to determine optimal execution strategies. In particular, given the set of all admissible strategies defined by

$$\mathcal{A} := \{(x_n)_{n=0}^N \,|\, (\mathcal{F}_n)_{n=0}^N \text{ - predictable}, x_0 = X, x_N = 0 \text{ and } x_{n-1} \geq x_n, \forall n = 1, ..., N\}, \tag{4.12}$$

we aim to find the optimal strategy $x^* = (x_n^*) \in \mathcal{A}$ that minimizes, for an initial state $v \in V$, the expected implementation shortfall:

$$\mathbb{E}_{0,v}[C(x)]. \tag{4.13}$$

Because we assume the return shocks ξ_n in equation (3.1) to be normally distributed with mean zero and independent of σ_n, we can reformulate the optimization problem as

$$\mathbb{E}_{0,v}[C(x)] = \mathbb{E}_{0,v}[Q_0(x)], \tag{4.14}$$

where

$$Q_n(x) := -\sum_{k=n+1}^{N} (x_{k-1} - x_k)^2 \frac{c}{2} + (x_{k-1} - x_k) \mathbb{E}_{k-1}^{\psi^{\gamma_{k-1} + \eta_k (x_{k-1} - x_k)}} [\sigma_k \xi_k]. \tag{4.15}$$

Given the set of all admissible strategies with a fixed position z for time t_k defined by

$$\mathcal{A}_k(z) := \{(x_n)_{n=0}^N \,|\, (\mathcal{F}_n)_{n=0}^N \text{ - predictable}, x_0 = X, x_k = z, x_N = 0 \text{ and } x_{n-1} \geq x_n, \forall n = 1, ..., N\}, \tag{4.16}$$

and stating the respective value function

$$J_n^v(z) := \min_{x \in \mathcal{A}_n(z)} \mathbb{E}_{0,v}[Q_n(x)], \tag{4.17}$$

we obtain the following result.

Theorem 4. *The value function J satisfies, for all $n = 1, ..., N-1$ and states $v \in V$, the Bellman equation*

$$J_{N-1}^v(x_{N-1}) = -\sum_{w \in V} p_{N-1}^{vw} \left(x_{N-1}^2 \frac{c}{2} + x_{N-1} \mathbb{E}_{N-1,v}^{\psi^{v\gamma + x_{N-1} w\eta}} [\sigma_N \xi_N] \right) \tag{4.18}$$

and

$$J_{n-1}^v(x_{n-1}) = \min_{0 \leq x_n \leq x_{n-1}} -\sum_{w \in V} p_{n-1}^{vw} \left((x_{n-1} - x_n)^2 \frac{c}{2} + (x_{n-1} - x_n) \mathbb{E}_{n-1,v}^{\psi^{v\gamma + (x_{n-1} - x_n) w\eta}} [\sigma_n \xi_n] \right.$$
$$\left. - J_n^w(x_n) \right). \tag{4.19}$$

The minimizing strategy $x_n^* \in \mathcal{A}_n(z)$, $n = 1, ..., N-1$, forms the optimal strategy for the risk-neutral objective.

Note that, depending on the distortion function and Markov chain specification, there may not be a unique solution.

Example 6 (Closed-form solution)**.** *For the assumptions of Example 4, we get a closed-form solution which corresponds to the constant speed strategy*

$$x_n^* = X \frac{N - n}{N} \tag{4.20}$$

as initially proposed by Bertsimas and Lo (1998). We refer to Theorem 3.1 and Remark 3.2 in Cheridito and Sepin (2014) for a partial proof.

4.2 Exponential objective

Starting with the work of Almgren and Chriss (2001), not only the expected implementation shortfall was taken into account, but also some measure of its variance. Here, we consider the exponential objective which, in some special case, gives the same results as the mean-variance objective used in Almgren and Chriss (2001). Hence, we aim to find the optimal strategy $x^* = (x_n^*) \in \mathcal{A}$ that minimizes, for an initial state $v \in V$ and a risk aversion parameter $\alpha > 0$, the expectation

$$\mathbb{E}_{0,v}[\exp\{\alpha C(x)\}]. \tag{4.21}$$

We again simplify the objective function by considering the equivalent

$$\mathbb{E}_{0,v}[\exp\{\alpha C(x)\}] = \mathbb{E}_{0,v}[\exp\{\alpha Q_0(x)\}], \tag{4.22}$$

where

$$Q_n(x) := -\sum_{k=n+1}^{N} (x_{k-1} - x_k)^2 \frac{c}{2} + (x_{k-1} - x_k)\mathbb{E}_{k-1}^{\psi^{\gamma_{k-1}+\eta_k(x_{k-1}-x_k)}}[\sigma_k \xi_k] + x_k \sigma_k \xi_k. \tag{4.23}$$

Then, we obtain for the value function, defined as

$$J_n^v(z) := \min_{x \in \mathcal{A}_n(z)} \mathbb{E}_{n,v}[\exp\{\alpha Q_n(x)\}], \tag{4.24}$$

the following result.

Theorem 5. *The value function J satisfies, for all $n = 1, ..., N-1$ and states $v \in V$, the Bellman equation*

$$J_{N-1}^v(x_{N-1}) = \sum_{w \in V} p_{N-1}^{vw} \exp\left\{-\alpha\left(x_{N-1}^2 \frac{c}{2} + x_{N-1}\mathbb{E}_{N-1,v}^{\psi^{v\gamma + x_{N-1}w\eta}}[\sigma_N \xi_N]\right)\right\} \tag{4.25}$$

and

$$J_{n-1}^v(x_{n-1}) = \min_{0 \le x_n \le x_{n-1}} \sum_{w \in V} p_{n-1}^{vw} \exp\left\{-\alpha\left((x_{n-1} - x_n)^2 \frac{c}{2} + (x_{n-1} - x_n)\mathbb{E}_{n-1,v}^{\psi^{v\gamma + (x_{n-1}-x_n)w\eta}}[\sigma_n \xi_n]\right)\right.$$
$$\left. + \frac{1}{2}\alpha^2 x_n^2 w_\sigma^2 \Delta t\right\} J_n^w(x_n). \tag{4.26}$$

The minimizing strategy $x_n^* \in \mathcal{A}_n(z)$, $n = 1, ..., N-1$, forms the optimal strategy for the exponential objective.

As in the risk-neutral case in Theorem 4, the optimal strategy might not be unique.

Example 7 (Closed-form solution)**.** *If we assume that σ and η are constant, $\gamma \equiv 0$ and $\Psi = \Psi^{Wang}$ (i.e., the setup in Example 4 with constant η), the strategy*

$$x_n^* = \frac{\sinh(\kappa(T - t_n))}{\sinh(\kappa T)} X, \tag{4.27}$$

for

$$\kappa = \frac{1}{\Delta t} \operatorname{arcosh}\left(1 + \frac{\alpha \sigma^2 \Delta t}{4\eta - 2c}\right) \tag{4.28}$$

is optimal, as derived in Almgren and Chriss (2001) for the mean-variance criterion $\mathbb{E}[C(x)] - \frac{\alpha}{2} Var(C(x))$. We refer to Remark 4.2 in Cheridito and Sepin (2014) for the proof.

5 Application and model analysis

We now elaborate on our modeling framework through a toy example, based on the parameters used in Cheridito and Sepin (2014) and Almgren and Chriss (2001). We additionally analyze how the new features of stochastic exogenous liquidity as well as different distortion functions from the Conic Finance framework impact the execution strategies.

Because in general there are no closed-form solutions for our optimization problems, we calculate the execution strategies numerically. As in Cheridito and Sepin (2014), we discretize the solution space by restraining trade sizes to multiples of 1% of the total position X. We first simulate 100,000 future paths of the Markov chain $(\sigma_n, \eta_n, \gamma_n)_{n=0}^N$. Then, for each simulation, we start at $x_0 = X$ the initial position and calculate the optimal position after the first trade x_1^*, or equivalently, the optimal trade size y_0^* for the time interval $(t_0, t_1]$. We do this by solving the backward equations defined in Theorems 4 and 5 for the risk-neutral and exponential objective functions, respectively. We then move forward one time step and calculate x_2^* in the same way, but now take into account the current realization $(\sigma_1, \eta_1, \gamma_1)$ and continue until x_{N-1}^* (as $x_N^* = 0$ by definition).

The distorted conditional expectations $\mathbb{E}_{n-1,v}^{\psi^{v_\gamma+(x_{n-1}-x_n)w_\eta}}[\sigma_n\xi_n]$ are thereby calculated via numerical integration as it can be shown that, for every $v, w \in V$ and $n = 1, ..., N$,

$$\mathbb{E}_{n-1,v}^{\psi^{v_\gamma+(x_{n-1}-x_n)w_\eta}}[\sigma_n\xi_n] = \int_0^\infty (1-\psi^{v_\gamma+(x_{n-1}-x_n)w_\eta}(F_{n-1,v}(x))dx - \int_{-\infty}^0 \psi^{v_\gamma+(x_{n-1}-x_n)w_\eta}(F_{n-1,v}(x))dx, \tag{5.1}$$

where $F_{n-1,v}(x)$ is the distribution of a normal mixture distributed random variable

$$Q_{n-1,v} \sim \sum_{w\in V} p_{n-1}^{vw} w_\sigma \mathcal{N}(0, \Delta t). \tag{5.2}$$

5.1 Toy example

To achieve results comparable to those reported in Cheridito and Sepin (2014), we use similar values for the parameters. In particular, we assume σ, η and γ to be independent and time-homogeneous. They can furthermore only take three different values, as defined in Table 1 and will always start in the low state. The transition probabilities for the processes are given by

$$p^\sigma = \begin{pmatrix} 0.9349 & 0.0434 & 0.0217 \\ 0.7164 & 0.2239 & 0.0597 \\ 0.44 & 0.48 & 0.08 \end{pmatrix}, \quad p^\eta = \begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0.15 & 0.8 & 0.05 \\ 0.05 & 0.05 & 0.9 \end{pmatrix}, \quad p^\gamma = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.7 & 0.1 \\ 0.2 & 0.5 & 0.3 \end{pmatrix}. \tag{5.3}$$

Since, as shown in Example 4, our model is an extension of the static volatility version of Cheridito and Sepin (2014), but with a slightly differently defined endogenous liquidity parameter η, we have rescaled the values for η reported in Cheridito and Sepin (2014).[46]

We also report the results for two other strategies, the *deterministic* and the *static* strategy. The deterministic strategy is the one specified in Examples 6 and 7. It is important to note that for this case, we do not need to impose the 1% trade size restriction for computational reasons as above. Hence, these strategies are inherently more efficient in that regard than the restricted ones. However, the deterministic strategy ignores the exogenous liquidity effect due to the process γ completely. For

[46] We divided the η values reported in Cheridito and Sepin (2014) by the mean over the three values $\sigma_{\text{low}}, \sigma_{\text{med}}, \sigma_{\text{high}}$. Rescaling using the steady state mean of σ resulted in optimal risk-neutral and exponential objective strategies that were very similar even for α values greater than 10^{-4}.

Parameter	Symbol	Value
Initial stock price	S_0	172
Initial position	X	35000
Duration	T	100 minutes
Number of subintervals	N	100
Permanent impact	c	2.5×10^{-7}
Volatility states	$\sigma_{\text{low}}, \sigma_{\text{med}}, \sigma_{\text{high}}$	3.51×10^{-3}, 3.3×10^{-2}, 1.172×10^{-1}
Endogenous liquidity states	$\eta_{\text{low}}, \eta_{\text{med}}, \eta_{\text{high}}$	1.95×10^{-5}, 9.76×10^{-5}, 4.88×10^{-4}
Exogenous liquidity states	$\gamma_{\text{low}}, \gamma_{\text{med}}, \gamma_{\text{high}}$	10^{-2}, 1.5×10^{-1}, 1
Distortion function	ψ	ψ_{Wang}
Risk aversion	α	10^{-5}

Table 1. Toy example parameter values.

the static case, we use the steady-state means of σ and η. The optimal execution is calculated according the backward equations of Theorems 4 and 5, but under the assumption that all processes are constant and equal to their respective steady-state means. In particular, the exogenous liquidity effect is taken into account, but the trade size restriction applies again.

Risk-neutral objective

When we impose a risk-neutral objective, the optimal execution strategies for the static and deterministic case naturally coincide, as there are 100 time steps and a 1% trade size restriction. In Figure 2, we plot the distribution of the optimal execution strategies and Figure 3 displays the corresponding implementation shortfall distribution.

From Table 2 we observe that the distribution of the implementation shortfall for the stochastic strategies has a significantly lower mean and median than for both the static and deterministic strategies and a slightly lower standard deviation.

Case	Mean	Median	Standard Deviation
Deterministic	332	318	4207
Static	332	318	4207
Stochastic	225	213	3947

Table 2. Descriptive statistics of the distribution of the implementation shortfall for the three cases of optimal execution strategies for the risk-neutral objective.

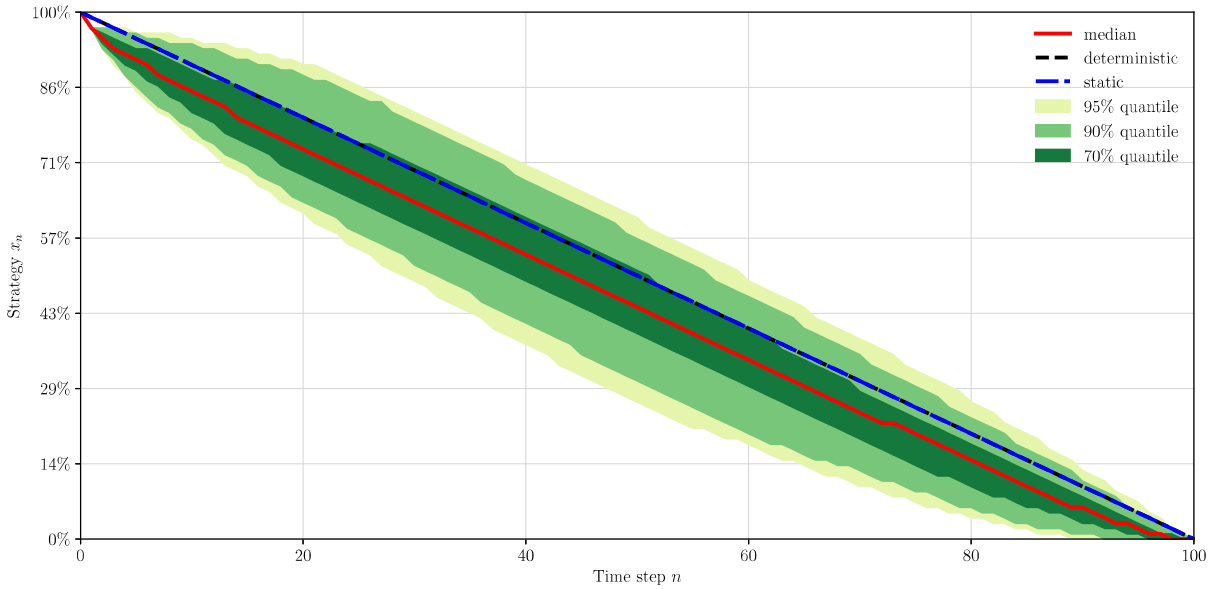


Figure 2. The distribution of the optimal execution strategies for the risk-neutral objective for the parameters described in Table 1. The yellow, light and dark green areas describe the 95th, 90th and 70th percentile of the distribution, respectively. The solid red line shows the median execution strategy, the dashed black line the deterministic strategy and the blue dash-dot line the static strategy.

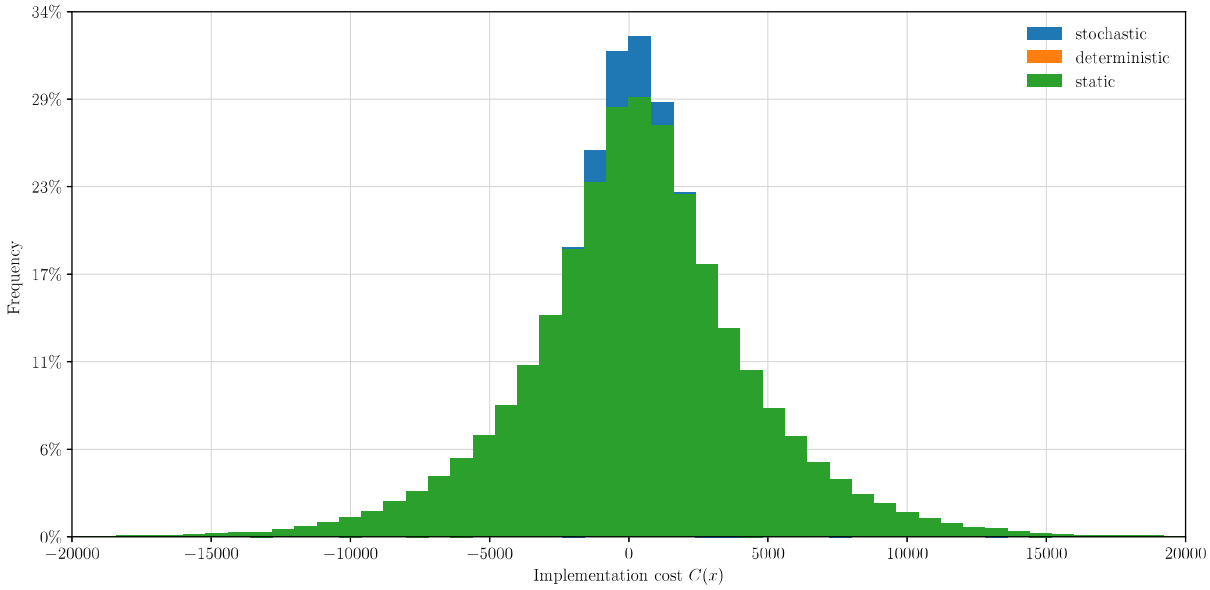


Figure 3. Histogram of the implementation shortfall of the optimal execution strategies for the risk-neutral objective as shown in Figure 2. The blue, orange and green areas describe the implementation shortfalls of the stochastic, deterministic and static strategies, respectively. The deterministic and static areas thereby overlap perfectly as the strategies are equivalent in this particular case.

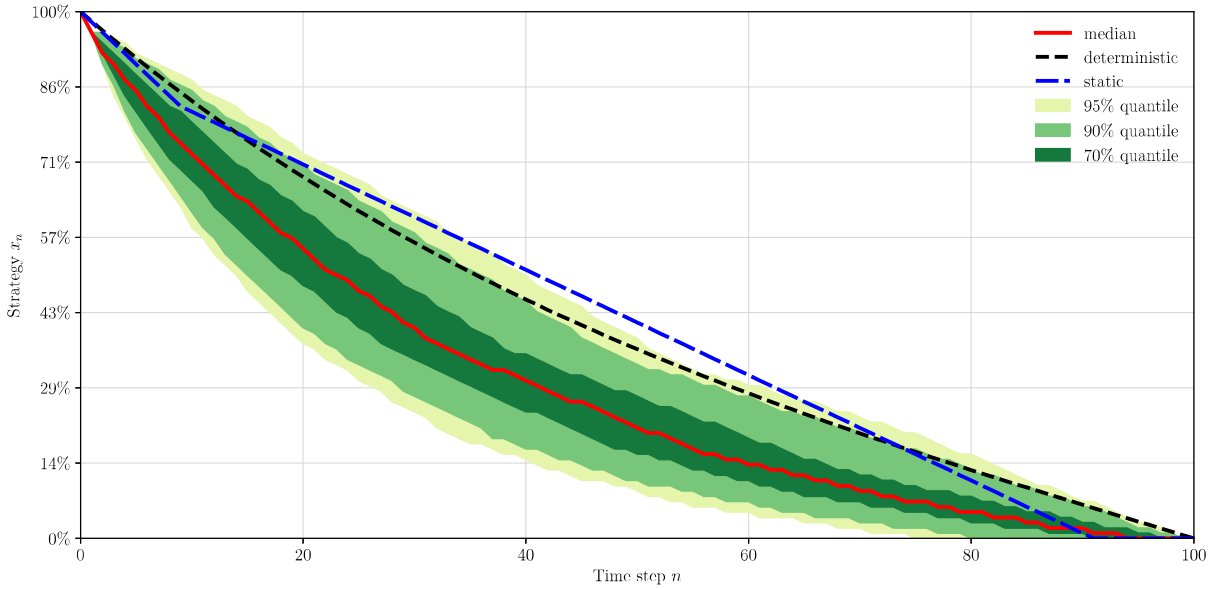


Figure 4. The distribution of the optimal execution strategies for the exponential objective for the parameters described in Table 1. The yellow, light and dark green areas describe the 95th, 90th and 70th percentile of the distribution, respectively. The solid red line shows the median execution strategy, the dashed black line the deterministic strategy and the blue dash-dot line the static strategy.

Exponential objective

In Figures 4 and 5 we plot the distribution of the optimal execution strategies and the corresponding implementation shortfall distribution. In line with our intuition, when we move from the risk-neutral to an exponential objective, then the optimal execution strategy is accelerated compared to the risk-neutral case.

We can further explore the effect of risk aversion by varying the risk aversion parameter α in the exponential objective. Figure 6 shows the mean and standard deviation of the exponential objective strategies over time for various degrees of risk aversion, $\alpha \in \{2 \times 10^{-6}, 10^{-5}, 2 \times 10^{-5}, 10^{-4}\}$. As expected, the higher the risk aversion, the faster the position is liquidated on average, increasing the standard deviation of the execution strategy at the beginning of the execution period and decreasing it towards the end. This finding is also consistent with the results of Almgren and Chriss (2001) and Cheridito and Sepin (2014).

As reported in Table 3, the implementation shortfall of the stochastic strategies have a significantly lower mean, median and standard deviation than both the static and deterministic strategy.

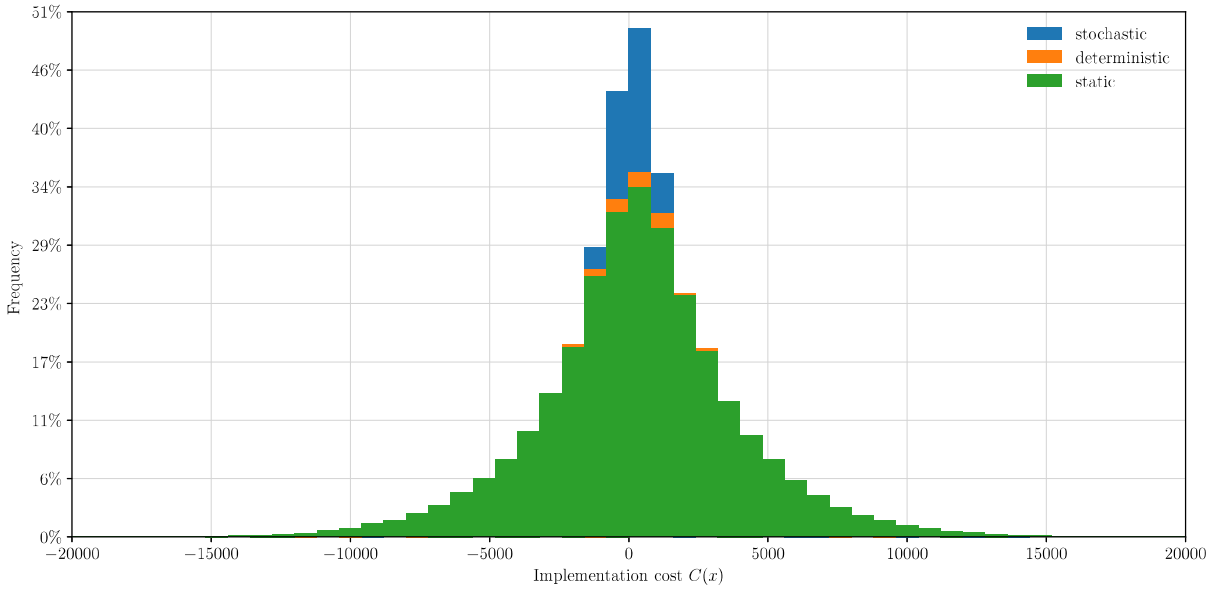


Figure 5. Histogram of the implementation shortfall of the optimal execution strategies for the exponential objective as shown in Figure 4. The blue, orange and green areas describe the implementation shortfalls of the stochastic, deterministic and static strategies, respectively.

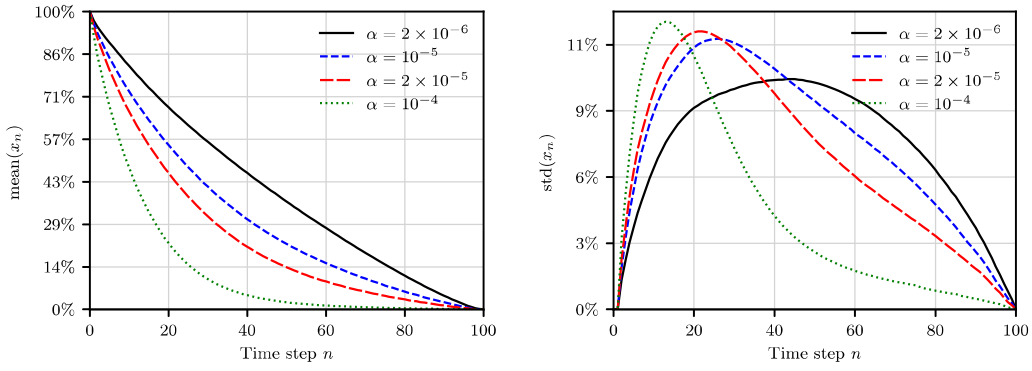


Figure 6. The mean (on the left) and standard deviation (on the right) of the optimal execution strategies for the exponential objective the parameters described in Table 1 but varying values of the risk-aversion parameter α.

It is worth noting that the static strategy has a slightly higher mean, median, and standard deviation than the deterministic strategy, which can mainly be attributed to the trade size restriction. Indeed, we find that rounding the trade amounts of the deterministic strategy to the nearest admissible size (which of course makes the strategy sub-optimal), results in a higher mean and median than the static strategy.

Case	Mean	Median	Standard Deviation
Deterministic	333	315	3567
Static	336	319	3694
Stochastic	233	227	2977

Table 3. Descriptive statistics of the distribution of the implementation shortfall for the three cases of optimal execution strategies for the exponential objective.

5.2 Deterministic exogenous liquidity effects

We now analyze the effect of the newly introduced exogenous liquidity parameter γ on the optimal execution strategies. To this end, we assume that the exogenous liquidity parameter is deterministic. We then calculate the optimal strategies for the same parameters and number of scenarios as in the previous section, but fix the deterministic exogenous liquidity process γ to either constant or time-varying.

Constant exogenous liquidity

For the case of constant exogenous liquidity, we start with no exogenous liquidity effect, i.e., $\gamma = 0$, and then set γ to 1 and 10. We summarize our results in Figure 7.

We observe that the optimal execution strategies for the risk-neutral objective barely change. A hint as to why the risk-neutral strategies are much less affected by the level of the constant exogenous liquidity parameter γ is provided by Example 3 as the optimal execution strategies under the risk-neutral objective function are independent on the level of γ, if constant, and if the volatility is constant as well.

In contrast to the execution strategy under the risk-neutral objective, the exponential strategies are significantly affected by changing the values of the exogenous liquidity parameter γ. Indeed, they

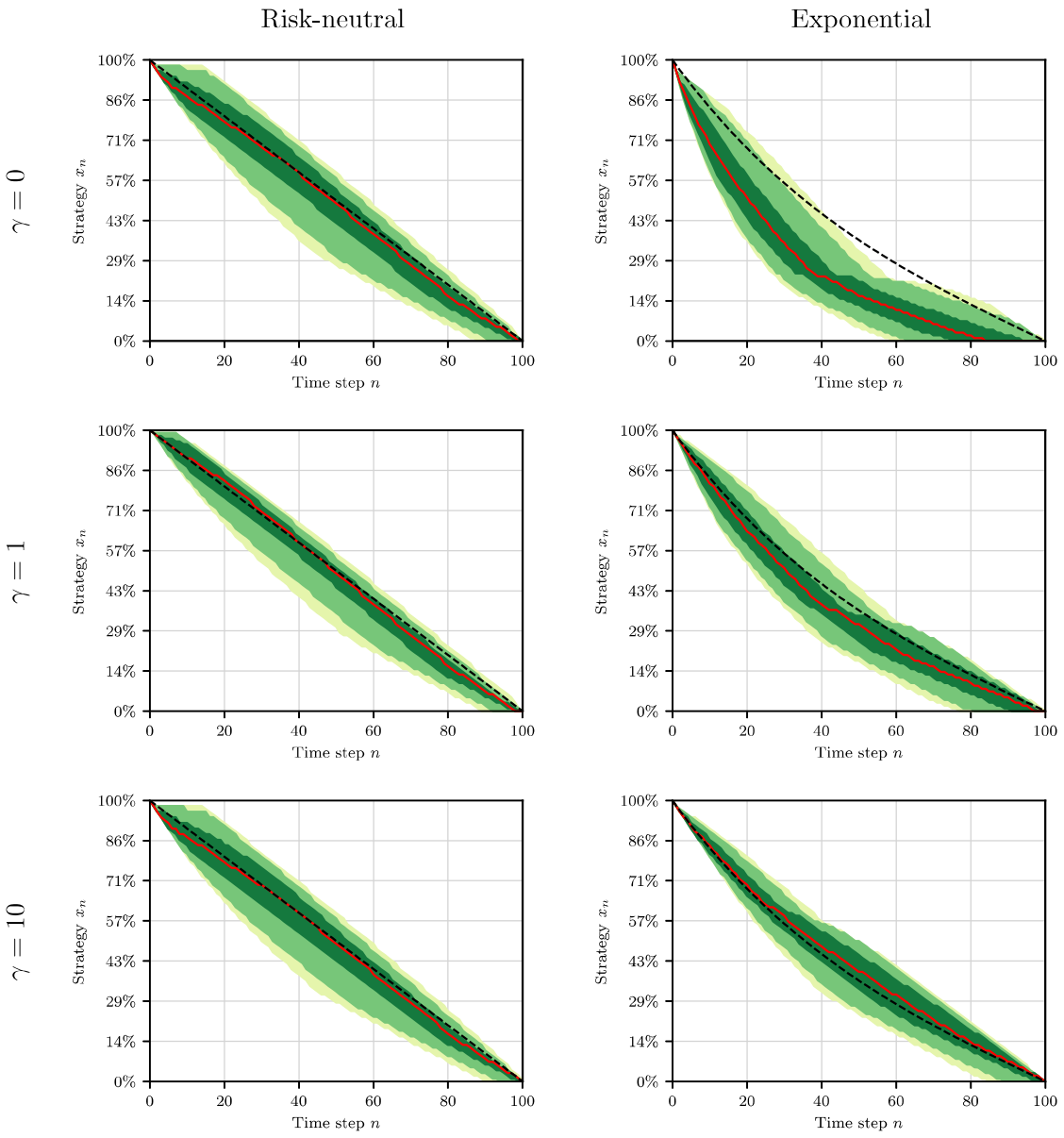


Figure 7. Comparison of the distributions of the optimal execution strategies for both objectives and varying degrees of constant exogenous liquidity γ. Left: Risk-neutral objective. Right: Exponential objective. First row: $\gamma = 0$. Second row: $\gamma = 1$. Third row: $\gamma = 10$. The other parameters are as specified in Table 1. The solid red line shows the median position size, the dashed black line the deterministic strategy.

lie almost up to the 95th quantile below the deterministic strategy for $\gamma = 0$. For $\gamma = 1$ the median becomes much closer to the deterministic strategy, and for $\gamma = 10$ it is even above it for most of the time steps.

Time-varying exogenous liquidity

To gain some further intuition, we now assume that exogenous liquidity is deterministic but time-varying. In particular, we replace the stochastic process γ_n as specified in Table 1 by three different time-varying but deterministic processes and leave everything else unchanged. We report the results in Figure 8.[47] For the first example of γ_n, which steadily increases from 0 to 1, the strategies for both objective functions liquidate the position faster than the deterministic strategies which do not take into account the exogenous liquidity cost would suggest.

In our second example for γ_n, which mimics the typical J-shape that we observe in Figure 1 for the IBM stock, the trade-off between selling according to the deterministic strategy and waiting for lower γ_n becomes obvious. Following a risk-neutral approach the trader would not sell anything at all until time step 40 and then increase the trading activity the closer to the end of the liquidation period he would get. On the other hand, according to the exponential objective, the trader would anyway have to sell a large part of his position over the first few time steps, but then slow down significantly and liquidate his remaining position as late as possible.

The last example, in which γ_n describes a wave starting and ending at 0.5 again shows how the two objectives can result in different trading behavior. The risk-neutral strategies sell more slowly than suggested by the deterministic strategy in the first 20 time steps, only to sell faster as γ_n decreases until $n = 80$. A large part of all strategies suggest to liquidate the whole position before γ_n increases again. The exponential objective strategies, however, try to sell as fast as possible to exploit γ_n being smaller than its maximum value in the first few time steps. The trading activity then decreases markedly and only picks up again as γ_n decreases.

[47] The exponential objective strategies for $\alpha = 10^{-5}$ look almost identical to the risk-neutral strategies, which is why we show the optimal exponential objective strategies for $\alpha = 10^{-4}$.

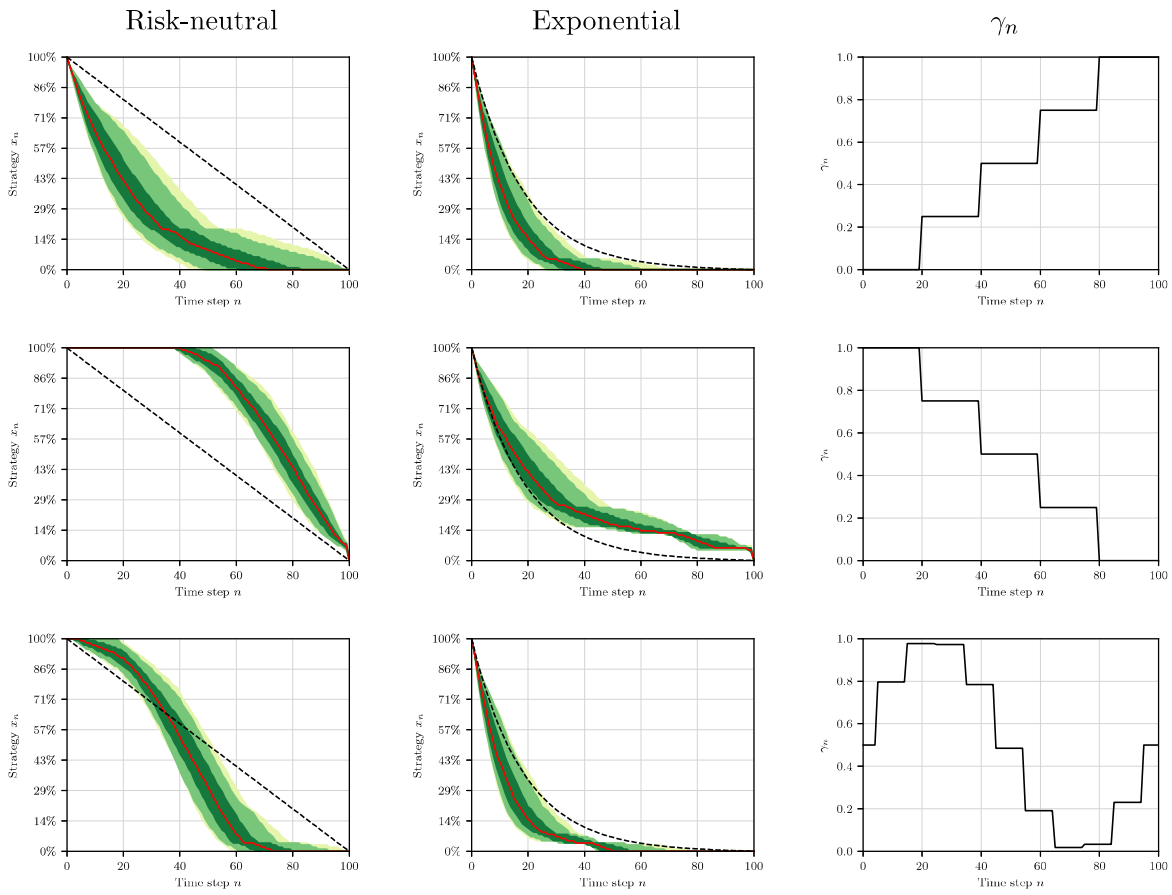


Figure 8. Comparison of the distributions of the optimal execution strategies for both objectives and various deterministic but time-varying processes (γ_n). Left: Risk-neutral objective. Middle: Exponential objective. Right: Process (γ_n). The risk-aversion parameter α for the exponential objective is set to 10^{-4}, the other parameters are as specified in Table 1. The solid red line shows the median position size, the dashed black line the deterministic strategy.

5.3 Distortion function effect

In the discussion above, we have focused on the Wang distortion function. This choice leads to a direct generalization of some models in the existing literature, as shown in Example 4. However, there is a multitude of distortion functions to choose from, a few of which have been mentioned previously. In Figure 9, we compare the optimal trading strategies based on the Wang and the CVaR distortion functions in Examples 1 and 2 based on the parameters of Table 1. As is evident on the top row, the optimal trading strategies do not change materially, if the Wang distortion (the results for which are plotted on the left) is exchanged by the CVaR distortion (as can be seen on the right). For the exponential strategies, however, the CVaR distortion results in a faster liquidation. The median position closely resembles the $\alpha = 2 \times 10^{-5}$ mean as shown in Figure 6, i.e., the CVaR distortion function results in a more risk-averse liquidation strategy. However, we must point out that the model parameters depend on the choice of distortion function and parameters calibrated for one may not be appropriate to be used with another. It would be an interesting avenue of future research to explore empirically different distortion functions that fit observed price behavior.[48]

6 Conclusion

We add to the existing literature on optimal execution in two ways. First, we use the recently developed theory of Conic Finance to provide a natural framework with which to model nonlinear temporary market impact. We show that our model extends classical optimal execution models such as the ones by Bertsimas and Lo (1998) and Almgren and Chriss (2001). This is the first time Conic Finance has been applied to the problem of optimal execution and in the process we extend the theory with distorted conditional expectations and random liquidity.

Second, we explicitly model exogenous liquidity which is due to other market participants as a stochastic process. This can be used to better take into account the known J-, U- or W-shapes in intraday bid-ask spreads when determining optimal execution strategies.

We derive the Bellman equations for the risk-neutral and exponential objective functions to de-

[48]Similarly, Bannör and Scherer (2014) calibrated a distortion function to fit observed end of day bid-ask prices.

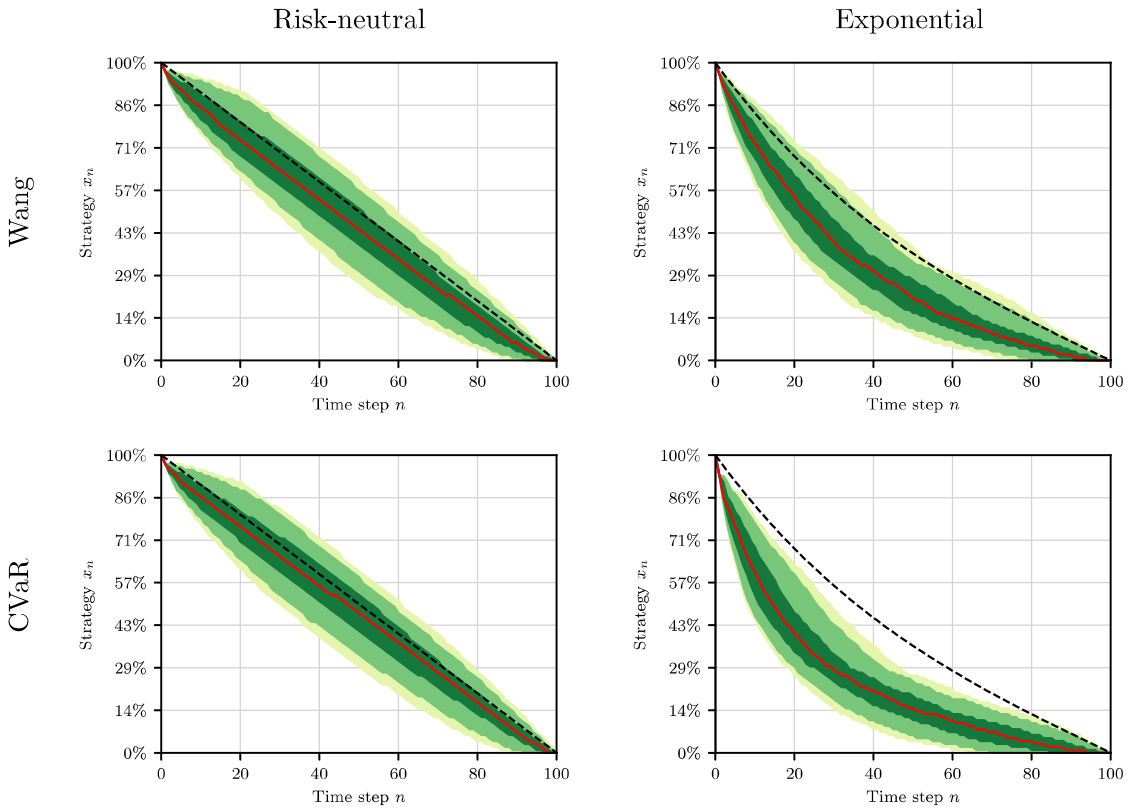


Figure 9. Comparison of the distributions of the optimal execution strategies for both objectives and two different distortion functions. Left: Risk-neutral objective. Right: Exponential objective. First row: Wang distortion function. Second row: CVaR distortion function. The other parameters are as specified in Table 1. The solid red line shows the median position size, the dashed black line the deterministic strategy.

termine optimal execution strategies and finally conclude by presenting a toy example and analyzing various aspects of our proposed model.

References

Albrecher, H., Guillaume, F., & Schoutens, W. (2013). Implied liquidity: Model sensitivity. *Journal of Empirical Finance*, *23*, 48–67.

Alfonsi, A., Fruth, A., & Schied, A. (2010). Optimal execution strategies in limit order books with general shape functions. *Quantitative Finance*, *10*(2), 143–157.

Alfonsi, A., Schied, A., & Slynko, A. (2012). Order book resilience, price manipulation, and the positive portfolio problem. *SIAM Journal on Financial Mathematics*, *3*(1), 511–533.

Almgren, R. (2003). Optimal execution with nonlinear impact functions and trading-enhanced risk. *Applied Mathematical Finance*, *10*(1), 1–18.

Almgren, R. (2012). Optimal trading with stochastic liquidity and volatility. *SIAM Journal on Financial Mathematics*, *3*(1), 163–181.

Almgren, R., & Chriss, N. (1999). Value under liquidation. *Risk*, *12*(12), 61–63.

Almgren, R., & Chriss, N. (2001). Optimal execution of portfolio transactions. *Journal of Risk*, *3*, 5–40.

Bannör, K. F., & Scherer, M. (2014). On the calibration of distortion risk measures to bid-ask prices. *Quantitative Finance*, *14*(7), 1217–1228.

Bertsimas, D., & Lo, A. W. (1998). Optimal control of execution costs. *Journal of Financial Markets*, *1*(1), 1–50.

Cheridito, P., & Sepin, T. (2014). Optimal trade execution under stochastic volatility and liquidity. *Applied Mathematical Finance*, *21*(4), 342–362.

Cherny, A., & Madan, D. B. (2009). New measures for performance evaluation. *Review of Financial Studies*, *22*(7), 2571–2606.

Choquet, G. (1953). Theory of capacities, In *Annales de l'institut fourier*.

Curato, G., Gatheral, J., & Lillo, F. (2016). Optimal execution with non-linear transient market impact. *Quantitative Finance*, 1–14.

Föllmer, H., & Schied, A. (2011). *Stochastic finance: An introduction in discrete time*. Walter de Gruyter.

Fukasawa, M., & Stadje, M. (2018). Perfect hedging under endogenous permanent market impacts. *Finance and Stochastics*, *22*(2), 417–442.

Gatheral, J. (2010). No-dynamic-arbitrage and market impact. *Quantitative Finance*, *10*(7), 749–759.

Gatheral, J., & Schied, A. (2013). Dynamical models for market impact and algorithms for optimal order execution. In J.-P. Fouque & J. Langsam (Eds.), *Handbook on systemic risk*. Cambridge University Press.

Gatheral, J., Schied, A., & Slynko, A. (2012). Transient linear price impact and Fredholm integral equations. *Mathematical Finance*, *22*(3), 445–474.

Guéant, O. (2013). Permanent market impact can be nonlinear. *arXiv preprint, arXiv:1305.0413*.

Guéant, O. (2016). *The financial mathematics of market liquidity: From optimal execution to market making* (Vol. 33). CRC Press.

Guéant, O., & Lehalle, C.-A. (2015). General intensity shapes in optimal liquidation. *Mathematical Finance*, *25*(3), 457–495.

Guéant, O., & Pu, J. (2015). Option pricing and hedging with execution costs and market impact. *Mathematical Finance*.

Huberman, G., & Stanzl, W. (2004). Price manipulation and quasi-arbitrage. *Econometrica*, *72*(4), 1247–1275.

Kratz, P., & Schöneborn, T. (2018). Optimal liquidation and adverse selection in dark pools. *Mathematical Finance*, *28*(1), 177–210.

Leippold, M., & Schärer, S. (2017). Discrete-time option pricing with stochastic liquidity. *Journal of Banking & Finance*, *75*, 1–16.

Lin, H.-Y., & Fahim, A. (2017). Optimal portfolio execution under time-varying liquidity constraints. *Applied Mathematical Finance*, 1–30.

Madan, D. B. (2010). Conserving capital by adjusting deltas for gamma in the presence of skewness. *Journal of Risk and Financial Management*, *3*(1), 1–25.

Madan, D. B., & Cherny, A. (2010a). Illiquid markets as a counterparty: An introduction to Conic Finance. *Robert H. Smith School Research Paper No. RHS*, 06–115.

Obizhaeva, A. A., & Wang, J. (2012). Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets*.

Schied, A., Schöneborn, T., & Tehranchi, M. (2010). Optimal basket liquidation for CARA investors is deterministic. *Applied Mathematical Finance*, *17*(6), 471–489.

Wang, S. S. (2000). A class of distortion operators for pricing financial and insurance risks. *Journal of Risk and Insurance*, *67*(1), 15–36.

Zwergel, B., & Heiden, S. (2014). Intraday futures patterns and volume–volatility relationships: The German evidence. *Review of Managerial Science*, *8*(1), 29–61.

Appendices

A Proofs

Proof of Equation 4.2. First, by the dynamic translation invariance property of conditional non-linear distorted expectations, the implementation shortfall is equal to

$$C(x) = XS_0 - \sum_{n=1}^{N} y_n \left(S_{n-1} + \mathbb{E}_{n-1}^{\psi^{\gamma_{n-1}+\eta_n y_n}} [\sigma_n \xi_n] \right). \tag{A.1}$$

Second, for $n = 1, ..., N$, it holds that

$$S_{n-1} = S_{n-2} + \sigma_{n-1}\xi_{n-1} - cy_{n-1} = \cdots = S_0 + \sum_{k=1}^{n-1} \sigma_k \xi_k - cy_k \tag{A.2}$$

which gives

$$\sum_{n=1}^{N} y_n S_{n-1} = XS_0 + \sum_{n=1}^{N} (x_{n-1} - x_n) \left(\sum_{k=1}^{n-1} \sigma_k \xi_k - c(x_{k-1} - x_k) \right) \tag{A.3}$$

as $\sum_{n=1}^{N} y_n = X$ by definition. By the fact that for any real-valued process $(z_n)_{n=0}^{N}$ we have

$$\sum_{n=1}^{N} (x_{n-1} - x_n) \sum_{k=1}^{n-1} z_k = \sum_{n=1}^{N} x_n z_n, \tag{A.4}$$

$$\sum_{n=1}^{N} x_n (x_{n-1} - x_n) = \frac{1}{2} x_0^2 - \frac{1}{2} \sum_{n=1}^{N} (x_n - x_{n-1})^2, \tag{A.5}$$

we can calculate

$$\sum_{n=1}^{N} (x_{n-1} - x_n) \left(\sum_{k=1}^{n-1} \sigma_k \xi_k - c(x_{k-1} - x_k) \right) = \sum_{n=1}^{N} x_n \sigma_n \xi_n - \frac{c}{2} X^2 + \frac{c}{2} \sum_{n=1}^{N} (x_n - x_{n-1})^2. \tag{A.6}$$

Combining (A.3) and (A.6) with (A.1) gives the simplified implementation shortfall

$$C(x) = \frac{c}{2} X^2 - \sum_{n=1}^{N} (x_{n-1} - x_n)^2 \frac{c}{2} + (x_{n-1} - x_n) \mathbb{E}_{n-1}^{\psi^{\gamma_{n-1}+\eta_n y_n}} [\sigma_n \xi_n] + x_n \sigma_n \xi_n. \tag{A.7}$$

□

A.1 Proof of Theorem 4

Proof. Since $x_N = 0$ we have

$$\begin{aligned} J^v_{N-1}(x_{N-1}) &= \mathbb{E}_{N-1,v}[Q_{N-1}(x)] = -\mathbb{E}_{N-1,v}\left[x^2_{N-1}\frac{c}{2} + x_{N-1}\mathbb{E}^{\psi^{\gamma_{N-1}+x_{N-1}\eta_N}}_{N-1}[\sigma_N\xi_N]\right] \\ &= -\sum_{w\in V} p^{vw}_{N-1}\left(x^2_{N-1}\frac{c}{2} + x_{N-1}\mathbb{E}^{\psi^{v\gamma+x_{N-1}w\eta}}_{N-1,v}[\sigma_N\xi_N]\right) \end{aligned} \tag{A.8}$$

Then, for $n = 1, ..., N$,

$$\begin{aligned} J^v_{n-1}(x_{n-1}) &= \min_{0\le x_n\le x_{n-1}} \mathbb{E}_{n-1,v}[Q_{n-1}(x)] \\ &= \min_{0\le x_n\le x_{n-1}} -\mathbb{E}_{n-1,v}\left[(x_{n-1}-x_n)^2\frac{c}{2} + (x_{n-1}-x_n)\mathbb{E}^{\psi^{\gamma_{n-1}+(x_{n-1}-x_n)\eta_N}}_{n-1}[\sigma_n\xi_n] + x_n\sigma_n\xi_n\right] \\ &\quad + \mathbb{E}_{n-1,v}[Q_n(x)] \\ &= \min_{0\le x_n\le x_{n-1}} -\sum_{w\in V} p^{vw}_{n-1}\left((x_{n-1}-x_n)^2\frac{c}{2} + (x_{n-1}-x_n)\mathbb{E}^{\psi^{v\gamma+(x_{n-1}-x_n)w\eta}}_{n-1,v}[\sigma_n\xi_n] - J^w_n(x_n)\right), \end{aligned} \tag{A.9}$$

where we drop the $x_n\sigma_n\xi_n$ term because $\mathbb{E}_{n-1,v}[\xi_n] = 0$ and ξ_n is independent of σ_n. Also, minimizing over $x \in \mathcal{A}_{n-1}(x_{n-1})$ is equivalent to minimizing over $x_n \in [0, x_{n-1}]$. □

A.2 Proof of Theorem 5

Proof. Since $x_N = 0$ we have

$$\begin{aligned} J^v_{N-1}(x_{N-1}) &= \mathbb{E}_{N-1,v}[\exp\{\alpha Q_{N-1}(x)\}] \\ &= -\mathbb{E}_{N-1,v}\left[\exp\left\{-\alpha\left(x^2_{N-1}\frac{c}{2} + x_{N-1}\mathbb{E}^{\psi^{\gamma_{N-1}+x_{N-1}\eta_N}}_{N-1}[\sigma_N\xi_N]\right)\right\}\right] \\ &= \sum_{w\in V} p^{vw}_{N-1}\exp\left\{-\alpha\left(x^2_{N-1}\frac{c}{2} + x_{N-1}\mathbb{E}^{\psi^{v\gamma+x_{N-1}w\eta}}_{N-1,v}[\sigma_N\xi_N]\right)\right\} \end{aligned} \tag{A.10}$$

Now note that $\mathbb{E}\left[e^{\xi}\right] = e^{\frac{1}{2}\Delta t}$ for $\xi \in \mathcal{N}(0, \Delta t)$. With that, for $n = 1, ..., N$,

$$\begin{aligned}
J^v_{n-1}(x_{n-1}) &= \min_{0\le x_n\le x_{n-1}} \mathbb{E}_{n-1,v}[\exp\{\alpha Q_{n-1}(x)\}] \\
&= \min_{0\le x_n\le x_{n-1}} \mathbb{E}_{n-1,v}\left[\exp\left\{-\alpha\left((x_{n-1}-x_n)^2\frac{c}{2} + (x_{n-1}-x_n)\mathbb{E}^{\psi^{\gamma_{n-1}+(x_{n-1}-x_n)\eta_N}}_{n-1}[\sigma_n\xi_n]\right)\right\}\right] \\
&\quad \times\ \mathbb{E}_{n-1,v}[\exp\{-\alpha x_n\sigma_n\xi_n\}]\,\mathbb{E}_{n-1,v}[\exp\{\alpha Q_n(x)\}] \\
&= \min_{0\le x_n\le x_{n-1}} \sum_{w\in V} p^{vw}_{n-1}\exp\left\{-\alpha\left((x_{n-1}-x_n)^2\frac{c}{2} + (x_{n-1}-x_n)\mathbb{E}^{\psi^{v\gamma+(x_{n-1}-x_n)w\eta}}_{n-1,v}[\sigma_n\xi_n]\right)\right. \\
&\qquad\qquad \left. +\ \frac{1}{2}\alpha^2 x_n^2 w_\sigma^2\Delta t\right\} J^w_n(x_n), \qquad \text{(A.11)}
\end{aligned}$$

where, again, minimizing over $x \in \mathcal{A}_{n-1}(x_{n-1})$ is equivalent to minimizing over $x_n \in [0, x_{n-1}]$. □

B Allowing for buying periods during liquidation

From the very beginning, and following previous literature such as Almgren (2012) and Cheridito and Sepin (2014), we have not allowed the trader to increase the position temporarily during the liquidation period. However, this restriction could hide some unwanted features of the model. In particular, while there is no dynamic arbitrage in our model as the permanent market impact function is linear[49], there could be transaction-triggered price manipulation strategies as defined by Alfonsi, Schied, et al. (2012)[50]. Arguably, the model used to define optimal execution strategies should not return predatory and potentially illegal strategies in which the fact that the trader controls a big part of the market is exploited.

B.1 Theoretical considerations

In the following, we will present how our proposed model can be extended to also allow for buying periods during the liquidation period. The extension to also allow for temporary selling when building up a stock position follows similarly. The proofs are omitted as they follow trivially from the original statements.

[49] Also see Gatheral (2010).

[50] I.e., the expected execution costs of the liquidation strategy can be lowered by intermediate buying periods.

We remove the constraint that $x_{n-1} \geq x_n$ for all $n = 1, ..., N$, but impose a maximum position that can be achieved, $x_{\max} \geq x_n$ for all $n = 0, ..., N$.[51] The implementation shortfall then changes to

$$\widetilde{C}(x) := XS_0 - \sum_{n=1}^{N} (x_{n-1} - x_n) \left(\mathbb{1}_{\{x_{n-1} \geq x_n\}} b_n^{\Psi,\gamma,\eta} + \mathbb{1}_{\{x_{n-1} < x_n\}} a_n^{\Psi,\gamma,\eta} \right). \tag{B.1}$$

As in Proposition 4.2, we can simplify this expression as follows.

Proposition 1 (Proposition 4.2 with buying periods)**.**

$$\begin{aligned} \widetilde{C}(x) = \frac{c}{2} X^2 - \sum_{n=1}^{N} \Big[& (x_{n-1} - x_n)^2 \frac{c}{2} + x_n \sigma_n \xi_n \\ & + (x_{n-1} - x_n) \Big(\mathbb{1}_{\{x_{n-1} \geq x_n\}} \mathbb{E}_{n-1}^{\psi^{\gamma_{n-1} + \eta_n (x_{n-1} - x_n)}} [\sigma_n \xi_n] \\ & + \mathbb{1}_{\{x_{n-1} < x_n\}} \mathbb{E}_{n-1}^{\psi^{\gamma_{n-1} + \eta_n \| x_{n-1} - x_n \|}} [\sigma_n \xi_n] \Big) \Big] \end{aligned} \tag{B.2}$$

Note that since $\xi_n \sim \mathcal{N}(0, \Delta t)$, $\mathbb{E}_{n-1}^{\psi^z}[-\sigma_n \xi_n] = \mathbb{E}_{n-1}^{\psi^z}[\sigma_n \xi_n]$ for random variables z.

The admissible strategies now also take into account that temporary buying is allowed.

Definition 21. *The set of all admissible strategies is given by*

$$\widetilde{\mathcal{A}} := \{ (x_n)_{n=0}^N \,|\, (\mathcal{F}_n)_{n=0}^N\text{-predictable}, x_0 = X, x_N = 0 \text{ and } x_n \leq x_{\max} \, \forall n = 0, ..., N \} \tag{B.3}$$

and the set of all admissible strategies with a fixed position z for time t_k is

$$\widetilde{\mathcal{A}}_k(z) := \{ (x_n)_{n=0}^N \,|\, (\mathcal{F}_n)_{n=0}^N\text{-predictable}, x_0 = X, x_k = z, x_N = 0 \text{ and } x_n \leq x_{\max} \, \forall n = 0, ..., N \}. \tag{B.4}$$

The dynamic programming equation then changes in the following way for the risk-neutral objective.

Theorem 6 (Theorem 4 with buying periods)**.** *The value function $\widetilde{J}$ satisfies, for all $n = 1, ..., N-1$ and states $v \in V$, the Bellman equation*

$$\widetilde{J}_{N-1}^v(x_{N-1}) = - \sum_{w \in V} p_{N-1}^{vw} \left(x_{N-1}^2 \frac{c}{2} + x_{N-1} \mathbb{E}_{N-1,v}^{\psi^{v\gamma + x_{N-1} w \eta}} [\sigma_N \xi_N] \right) \tag{B.5}$$

[51] This maximum constraint could, at least theoretically, also be $+\infty$.

and

$$\begin{aligned}
&\widetilde{J}^v_{n-1}(x_{n-1}) \\
&= \min_{0\le x_n\le x_{\max}} -\sum_{w\in V} p^{vw}_{n-1}\Big((x_{n-1}-x_n)^2\frac{c}{2} \\
&\qquad + (x_{n-1}-x_n)\Big(\mathbb{1}_{\{x_{n-1}\ge x_n\}}\mathbb{E}^{\psi^{v\gamma+(x_{n-1}-x_n)w\eta}}_{n-1,v}[\sigma_n\xi_n] \\
&\qquad + \mathbb{1}_{\{x_{n-1}<x_n\}}\,\mathbb{E}^{\psi^{v\gamma+\|x_{n-1}-x_n\|w\eta}}_{n-1,v}[\sigma_n\xi_n]\Big) - \widetilde{J}^w_n(x_n)\Big).
\end{aligned} \tag{B.6}$$

The minimizing (x_n^*) *form the optimal strategy for the risk-neutral objective.*

For the exponential objective, the dynamic programming equation is given by the following Theorem.

Theorem 7 (Theorem 5 with buying periods)**.** *The value function* $\widetilde{J}$ *satisfies, for all* $n = 1, ..., N-1$ *and states* $v \in V$*, the Bellman equation*

$$\widetilde{J}^v_{N-1}(x_{N-1}) = \sum_{w\in V} p^{vw}_{N-1}\exp\left\{-\alpha\left(x^2_{N-1}\frac{c}{2} + x_{N-1}\mathbb{E}^{\psi^{v\gamma+x_{N-1}w\eta}}_{N-1,v}[\sigma_N\xi_N]\right)\right\} \tag{B.7}$$

and

$$\begin{aligned}
\widetilde{J}^v_{n-1}(x_{n-1}) = \min_{0\le x_n\le x_{\max}} \sum_{w\in V} p^{vw}_{n-1}\exp\Big\{-\alpha\Big((x_{n-1}-x_n)^2\frac{c}{2} \\
+ (x_{n-1}-x_n)\Big(\mathbb{1}_{\{x_{n-1}\ge x_n\}}\mathbb{E}^{\psi^{v\gamma+(x_{n-1}-x_n)w\eta}}_{n-1,v}[\sigma_n\xi_n] \\
+\mathbb{1}_{\{x_{n-1}<x_n\}}\mathbb{E}^{\psi^{v\gamma+\|x_{n-1}-x_n\|w\eta}}_{n-1,v}[\sigma_n\xi_n]\Big) \\
+\frac{1}{2}\alpha^2x_n^2w_\sigma^2\Delta t\Big\}\,J^w_n(x_n).
\end{aligned} \tag{B.8}$$

The minimizing (x_n^*) *form the optimal strategy for the exponential objective.*

B.2 Empirical evidence and conclusion

We have re-ran all examples presented in this paper, and a few others, with $x_{\max} = 2X$, but have not found a single case of transaction-triggered price manipulation. This is of course not to say that these might not exist, e.g., for $x_{\max} > 2X$, when allowing for trades of all sizes or for different parameter

combinations. However, it does give an indication that there is no systematic problem with these kinds of predatory execution strategies. Furthermore, we would argue along the lines of Curato et al. (2016) that price manipulation strategies are much more susceptible to model misspecifications. And should our proposed model result in transaction-triggered price manipulations, we would suggest to use the process γ_n as a regularization term, similarly as proposed by Curato et al. (2016) who could successfully prevent their model from choosing execution strategies with intermediate buying periods by adding spread costs.

Deep Learning for (Intra-Horizon) Value-at-Risk Forecasting and Application to Conditional Value-at-Risk

Steven Schärer

Abstract

We propose a deep neural network model to forecast Value-at-Risk and Intra-Horizon Value-at-Risk over multiple time horizons and quantile levels. We demonstrate with a large-scale experiment involving 500 current and former constituents of the S&P 500 that this model outperforms other models typically suggested in the literature and used in practice with regards to the quantile regression loss function. Additionally, once trained our model does not require frequent re-training and can be used with less data than is often required for alternative models. We finally motivate how the proposed model can be extended to forecast Conditional Value-at-Risk which we demonstrate in an example.

JEL Classification: G17, G32, C45, C51, C52, C53

Keywords: Deep Learning, LSTM, Quantile Regression, Value-at-Risk, Intra-Horizon Value-at-Risk

1 Introduction

Adequately estimating or forecasting measures of risks of financial positions is a problem that is encountered and at times also regulated in many areas in finance. Examples include capital requirement calculations for insurance companies, initial margin calculations for central counterparties and risk management in wealth management with time horizons ranging from days to months.[52] In many cases, it is common to use Value-at-Risk (VaR) or Conditional Value-at-Risk (CVaR) as a measure of risk.[53] Despite some well-known shortcomings and a slow regulatory shift to CVaR, we will concentrate on VaR in this paper as it is still more widely used. We will, however, demonstrate in Section 4 that the proposed model can be used to forecast CVaR and motivate how the model can be extended if forecasting CVaR is one of its main purposes.

Historically, VaR has first been calculated non-parametrically as the quantile of a history of asset returns. For example, a 5-day 95% VaR would be calculated as the 5% quantile of some history of 5-day returns. This does not need any other assumption besides the stationary of the returns.[54] However, testing this way of calculating VaR out of sample has generally shown unsatisfactory results, with estimates that are slow to react to changing market conditions, being too low when volatility rises but also too high in times of prolonged tranquility as can be seen in an example in Figure 1a. Another issue is the fixed window of returns that is usually used which, if a period of large volatility drops out of the window, can cause drastic changes in the VaR estimate in a short period of time. In the case of margin computations at a central counterparty this can cause widespread changes in margin requirements without stock prices necessarily moving by a lot.

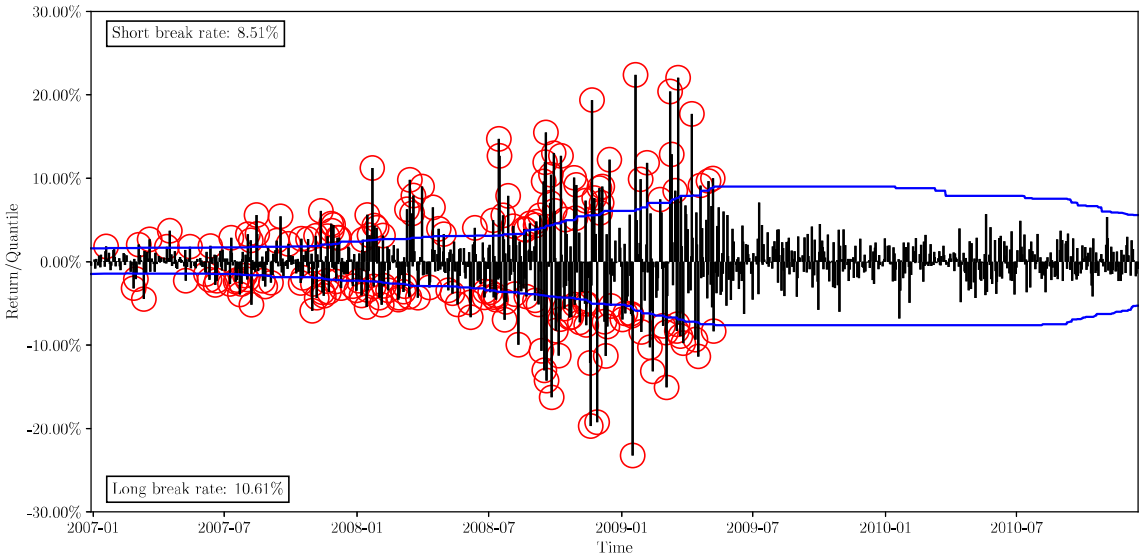
The next innovation in the calculation of VaR came with the RiskMetrics™ methodology developed by J.P. Morgan in the early 1990s which assumes a parametric model of the underlying returns[55]. This model featured a much greater reactivity and provided better out-of-sample results, also see

[52] Relevant regulations are for example Solvency II, Basel II, EMIR and MiFID II.

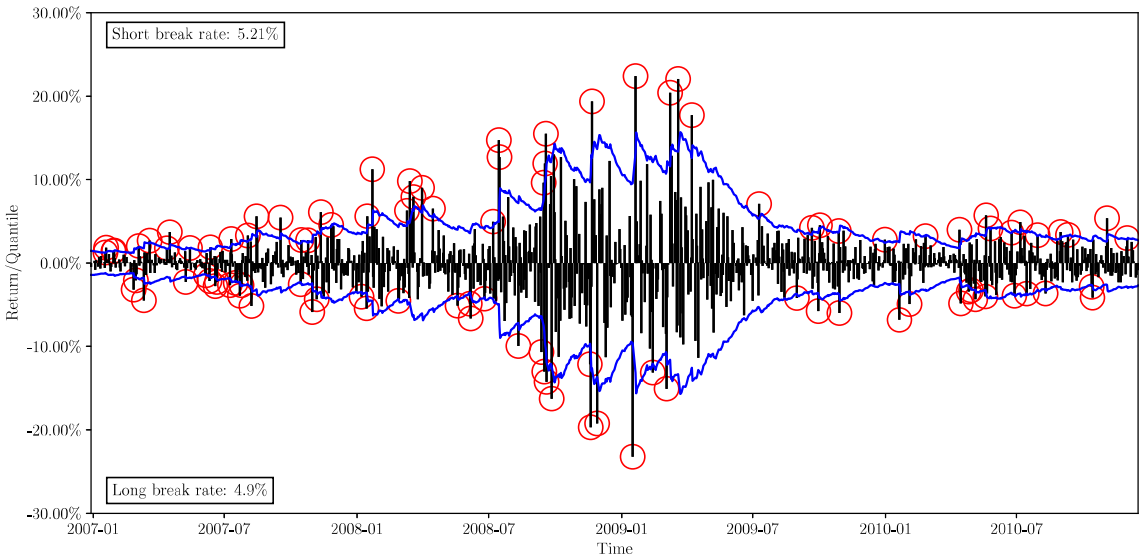
[53] See, e.g., Artzner, Delbaen, Eber, and Heath (1999) and Föllmer and Schied (2011).

[54] Alternatively and more frequently, a 1-day VaR is calculated and then multiplied by $\sqrt{5}$ according to the square-root of time rule. This requires much stronger assumptions, such as the independence of returns. More complicated scaling rules can be derived if these assumptions are weakened.

[55] It is assumed that the volatility follows a EWMA process and that the residuals are standard-normally distributed. See Appendix A for more details.



(a) Hist



(b) EWMA

Figure 1. We plot the out-of-sample 5% and 95% 1-day ahead VaR forecasts for JPM from 2007-01-01 to 2010-12-31. The top model is Hist with a lag of 500, the bottom model is EWMA with a lag of 500, $\lambda = 0.94$, mean zero and assuming normal innovations (see Appendix A for more information). Red circles indicate violations of the VaR forecast. The figure shows that Hist reacts very slowly to an ever increasing number of violations due to higher volatility. After volatility decreases the quantile estimates remain elevated for a prolonged period of time, resulting in no violations at all instead of the 5% target. EWMA, on the other hand, is able to react fairly quickly to spikes in volatility after which it returns a bit more slowly to levels appropriate for the current market conditions. It is able to produce the target of 5% violations in both directions fairly accurately over the displayed time period.

Figure 1b for an example. Indeed, similar models are still in use today and there is on-going research into making the modeling assumptions more realistic.[56]

As already mentioned, the most well-known alternative to VaR that has been proposed in the literature is CVaR and deals with the drawback of VaR giving no information about the tail of the distribution beyond the VaR level. In a multi-period setting, the fact that VaR only measures the risk of the distribution at the end of the horizon is another possible drawback that has attracted much less attention. This can be appropriate if the aim is to measure the risk of, e.g., a buy-and-hold position. But in environments such as trading desks or central counterparties where there are possible margin calls if the value of a position goes below a certain level this can be a significant disadvantage. For that reason, intra-horizon Value-at-Risk (iVaR)[57] has been suggested as an alternative to VaR. It is defined as a quantile of the distribution of the minimum cumulative return over the whole horizon, compared to a quantile of the cumulative return distribution at the end of the horizon for VaR.[58] IVaR has first been suggested in Boudoukh et al. (2004) and was later extended to a GARCH-type model in Bhattacharyya et al. (2009) and to jump models in Bakshi and Panayotov (2010) and Leippold and Vasiljevic (2018). We will consider both VaR and iVaR in this paper and will demonstrate that our proposed model can be trained to forecast both risk measures.

What most of the previously mentioned models have in common is that they stipulate rather restrictive dynamics on the movements of stock prices, which, when studied empirically, often are not satisfied. This has led to the creation of ever more complex models and distributional assumptions. However, while more complex models are designed to fit in-sample data better, the out-of-sample performance does not always improve. This can either be caused by overfitting, or because the more complex models require ever more data for a robust calibration. The latter can be a significant issue, forcing practitioners to use proxies or other workarounds until enough data is available.

Instead of presenting yet another GARCH extension, we propose a model that does not impose a particular data generation process but rather learns how to predict VaR and iVaR purely based on

[56]Popular models are the updated RiskMetrics™ methodology Zumbach (2007), semi-parametric versions that resample the historical residuals instead of assuming a distribution (see, e.g., Barone-Adesi et al. (1998)) or more flexible alternatives to the EWMA process and the normal distribution. See Appendix A for more details.

[57]IVaR is also known as MaxVaR.

[58]Or the distribution of the maximum cumulative return over the whole horizon, if the right-hand side of the return distribution is of interest.

historical data. This effort is supported by the steadily increasing availability of computational power and the recent explosion in the research of machine learning and especially deep learning techniques. In particular, we propose a recurring neural network (RNN) structure to forecast quantiles, based on a loss function provided by the theory of quantile regression.[59] The key building block is a long short-term memory (LSTM) layer, an RNN variant introduced in the seminal paper by Hochreiter and Schmidhuber (1997) which has become extraordinarily popular in recent years for various applications including machine translation, image captioning and time series forecasting. Very superficially, an LSTM layer recursively encodes a sequence of lagged data into a set of hidden states which can then, for example, be transformed into predictions of future data.

A feature of deep learning models is the potentially very high number of parameters that need to be trained, which generally requires a correspondingly large set of data for a robust training. This can be an issue when applying these techniques to financial markets as the history of end-of-day stock prices is much lower than the millions of data points used in other areas where deep learning models are applied. We work around this issue in multiple ways. First, our model predicts risk figures for multiple quantile levels and time horizons, all sharing the same LSTM layer. This increases the number of targets by a factor of 20 in the setup of our experiments. Second, we claim that there are universal features to the movements of stock prices and aim to learn those, rather than develop stock-specific models. To that end, we train the model on 250 stocks, without providing it any identifying information. This increases the number of available data points by a factor of 250, with the additional benefit of decreasing the chance of the model not being able to deal with a regime change or other behavior previously not encountered for a specific stock. In our experiments we train the model with daily data over a period of twenty years with a burn in-period of at most 500 data points. In total, this results in approximately 10 million targets that the model is calibrated against. We train the model just once and also apply it to instruments not encountered during training, therefore just requiring at most 500 data points to forecast risk measures. This is on the lower end of how much data is usually used to calibrate the more complex GARCH-like models described in the literature.

The first time a quantile prediction model using deep learning techniques has been suggested

[59] See Koenker and Bassett Jr (1978) for more information about quantile regression.

in the literature was by Taylor (2000), although besides the returns themselves he also used the one-step-ahead volatility forecasts of a GARCH model as inputs. Xu et al. (2016) present a neural network model inspired by the CAViaR model of R. F. Engle and Manganelli (2004), though only for single instruments, horizons and quantile levels. A model for predicting quantiles for multiple horizons and quantile levels has first been suggested by Wen et al. (2017) but not specifically with the intent of applying it to VaR or iVaR forecasting. Yan et al. (2018) present a model to predict the whole return distribution on a single horizon and for single instruments with a similar approach as ours, but then also impose a functional form of the quantile function. On the one hand, this removes the possibility of quantile predictions crossing each other (i.e., the α-quantile being smaller than the β-quantile even though $\alpha > \beta$). On the other hand, this again introduces a restriction, though somewhat flexible in their proposal, on the output space of the model. Sirignano and Cont (2019) have also suggested pooling a large number of instruments to learn universal features of financial markets and use it to predict high-frequency price movements on the order book level.

The rest of this paper is organized as follows. In Section 2, we present our proposed model and describe the results of various experiments in Section 3. An extension of the model to forecasting CVaR is presented in Section 4. Our findings are summarized in Section 5. The appendix contains the descriptions of the various benchmark models (Appendix A) and the statistical tests (Appendix B) typically used to determine the goodness of VaR models.

2 Proposed model

Assume we are at time t. We propose a deep neural network, called LSTM-QR, to forecast quantiles z^q_{t+k} for levels $q \in (0, 1)$ and time horizons $k >= 1$ based on a set of l lagged 1-day returns $x_{t-l+1}, ..., x_t$. In a risk management context these quantiles will then be used as k-day ahead VaR or iVaR estimates at level $1-q$, i.e., they estimate the quantiles of either the end-of-horizon distribution, or the quantiles of the distribution of the minimum, respectively maximum cumulative returns over horizon k, depending on q.[60]

[60] For $k = 1$, iVaR and VaR naturally coincide.

2.1 Network architecture

Consider the architecture described in Figure 2. A layer of LSTM units is used to encode l lagged feature vectors $f_{t-l}, ..., f_t$ into n hidden states h, i.e.,

$$(h_1, ..., h_n) = \text{LSTM}(f_{t-l}, ..., f_t). \tag{2.1}$$

These hidden states are then transformed to k-period quantile predictions via a densely-connected layer with a linear activation function for each horizon and quantile level, i.e., for all k and q

$$z^q_{t+k} = \sum_{j=1,...,n} w_j h_j + b, \tag{2.2}$$

where b is a bias term coming from the LSTM layer. As such, we are using a single LSTM layer to encode all relevant information to forecast quantiles at various levels and horizons. This allows for a more efficient use of computational time as opposed to using a single LSTM layer per quantile of interest. It also enables the model to learn the LSTM parameters, which are much larger in number than the parameters of the densely-connected layers, from various quantile levels and horizons, thereby improving training even further.

Due to the recursive nature of LSTM, an alternative way of modeling multiple future steps with LSTM would be to predict one step ahead, feed that forecast back into LSTM by treating it as a new data point to generate a two steps ahead prediction and so on. This is used for example in Natural Language Processing using Sequence-to-Sequence models. However, this recursive approach accumulates errors over the different forecasting steps and various studies have shown that directly targeting multi-step ahead predictions may be preferable in many circumstances, see, e.g., Taieb and Atiya (2015).

We have tested allowing for stacked LSTM layers to enable the model to learn more complex behavior as suggested in, e.g., Pascanu et al. (2013). While, as expected, increasing the number of layers while keeping all other parameters fixed does result in a better fit on the training set, the performance on the validation set tended to be slightly worse when using more layers. We have also considered using Multilayer Perceptrons (MLPs) instead of simple densely-connected layers after

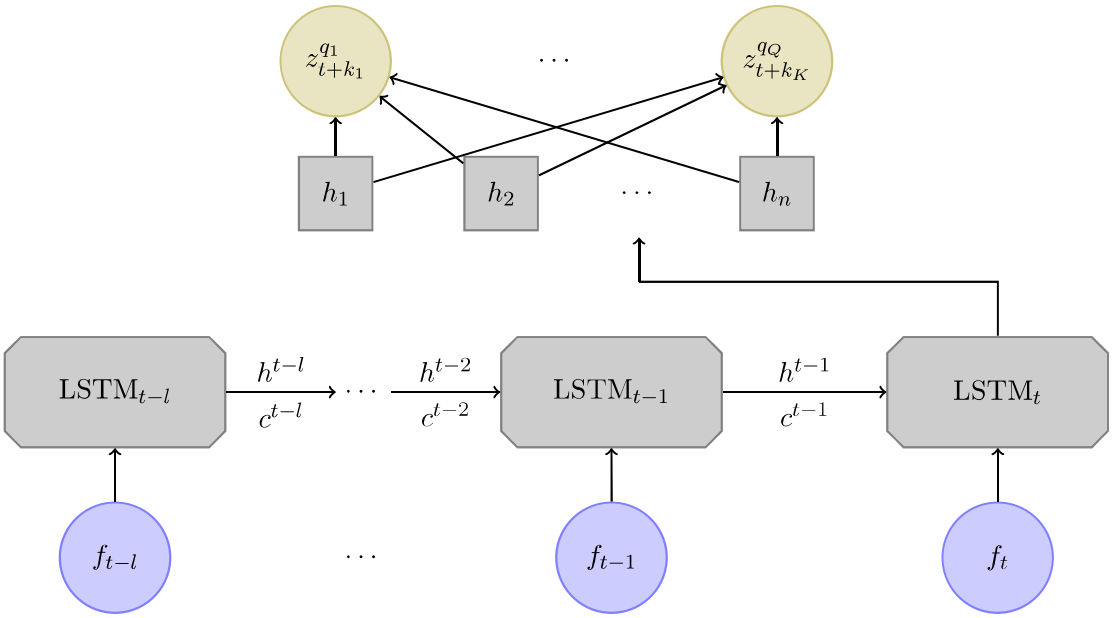


Figure 2. LSTM-QR network architecture for forecasting quantiles $z^{q_j}_{t+k_i}$ for $i = 1, ..., K$ and $j = 1, ..., Q$ based on features $f_{t-l}, ..., f_t$. The LSTM unit LSTM$_s$ takes as input a vector of features f_s as well as the memory c^{s-1} and hidden state vector h^{s-1} from the previous unit, if applicable. The hidden state vector of the last unit, $h = h^t$, is transformed into outputs $z^{q_j}_{t+k_i}$ via a densely connected network for each output.

applying the LSTM layer as in Wen et al. (2017), but also did not find it to improve the performance of the model. As such we opt for the simplest model, which reduces the number of parameters that have to be calibrated.

While computing the hidden states, LSTM units combine information from previous units with new information in a fashion that resembles the exponential smoothing that is applied in GARCH-like models. For more details on the internal workings of LSTM units we refer to Hochreiter and Schmidhuber (1997). We use a hyperbolic tangent activation function for the cell and hidden state and a sigmoid activation function for the input, forget and output gates.

2.2 Features

The first feature vector consists of the lagged returns $x_{t-l+1}, ..., x_t$. Empirical evidence and in particular the success of GARCH-like models in forecasting quantiles of financial return time series show that besides the returns themselves, the moments thereof can also contain valuable information. While it would be conceivable to let our model learn by itself that it is beneficial to also consider

the higher moments of the data points it encounters, it is more efficient to provide them directly as additional feature vectors. Therefore, and as Yan et al. (2018), we also add the second to fourth moments of the lagged returns to the set of features $f = (f_i)_{i=1,...,l}$,

$$f_i = \left(x_{t-i+1}, (x_{t-i+1} - \bar{x}_t)^2, (x_{t-i+1} - \bar{x}_t)^3, (x_{t-i+1} - \bar{x}_t)^4\right) \top \in \mathbb{R}^4 \tag{2.3}$$

with $\bar{x}_t := \frac{1}{l} \sum_{i=1}^{l} x_{t-i+1}$.

Other possible features could include reference data such as sectors or ratings, or future information such as seasonality, where applicable. The static attributes could be repeated over time and added to the set of feature vectors. Also see Wen et al. (2017) for a proposal how static and future information can be treated within the same model as time series data. Investigating whether the model performance improves with such additional features is left for future research.

2.3 Loss function

Assume there is a set of daily returns $x_1, ..., x_T$ on which the model should be trained or evaluated, with $T \geq l + k$. The model takes l data points as input, $x_{t-l}, ..., x_t$, and outputs quantiles z^q_{t+k}. Quantiles, however, are not directly observable from the data, so we cannot use traditional methods of judging the goodness of the forecasts such as mean squared error. Instead, we only have the actual return realizations, $y = x_{t+k} = x_t + ... + x_{t+k}$ for the case of VaR or $y = x_{t+\tau(q)}$, where

$$\tau(q) = \begin{cases} \arg\min_{j=1,...,k} x_{t+j} & \text{if } q \leq 0.5 \\ \arg\max_{j=1,...,k} x_{t+j} & \text{otherwise} \end{cases} \tag{2.4}$$

for iVaR. The theory of quantile regression[61] provides a way to measure the goodness of quantile forecasts with respect to future realizations via the loss function $L_q(y, z)$ defined by

$$L_q(y, z) = \begin{cases} q(y - z) & \text{if } y - z \geq 0 \\ (q - 1)(y - z) & \text{otherwise} \end{cases}.$$

[61] See Koenker and Bassett Jr (1978).

Since our model returns quantile estimates for multiple horizons and quantile levels, and is evaluated over all available time steps, the model is evaluated with respect to the goal of minimizing the total loss,

$$L^{\text{total}}(x, z) = \sum_k \frac{1}{T-k-l+1} m^{\text{horizon}}(k) \sum_q m^{\text{level}}(q) \sum_{t=l}^{T-k} L_q(x_{t+k}, z^q_{t+k}). \tag{2.5}$$

We thereby apply two different multipliers for the various time horizons and quantile levels, to ensure that errors on different horizons and quantile levels count approximately the same towards the overall error. First, the multiplier over the horizon dimension,

$$m^{\text{horizon}}(k) := \frac{1}{\sqrt{k}} \tag{2.6}$$

is simply the square-root of time scaling rule.[62] Second, for the multiplier for different quantile levels consider that we are only going to be interested in levels $\{0.01, 0.05, 0.95, 0.99\}$. For those levels we propose to use the multiplier

$$m^{\text{level}}(q) = \begin{cases} 3 & \text{if } q \in \{0.01, 0.99\} \\ 1 & \text{otherwise} \end{cases} \tag{2.7}$$

and refer to Appendix C for theoretical and empirical justifications for choosing this multiplier. The arguments contained therein can be extended to other quantile levels without loss of generality.

While the possibility of using different multipliers for different time horizons and quantile levels is mentioned in Wen et al. (2017), no specific choice is suggested for the problem we are interested in. Other approaches suggested in the literature either only model single quantile levels and horizons (Xu et al. (2016)), or do not consider a modified weighting scheme (Yan et al. (2018)).

Finally, when training or validating the model we actually use data of multiple instruments $i = 1, ..., N$ in our training set. The total loss function of the training set is therefore defined as

$$L^{\text{total,train}}(x, z) = \frac{1}{N} \sum_i L_t^{\text{total}}(x^{(i)}, z^{(i)}), \tag{2.8}$$

[62] While the ubiquitous and often unquestioning use of the square-root of time scaling rule in finance is often validly criticized, in this and many other contexts it can functions as an acceptable approximation. Also see Brummelhuis and Kaufmann (2004)

where superscript (i) denotes the returns and quantile forecasts specific to instrument i.

3 Experiments

We test whether by training LSTM-QR once on a large set of instruments the model outperforms a set of benchmark models, as described in Appendix A, on both large out-of-sample time periods and instruments not part of the training set. In particular, we will investigate the following hypotheses:

(i) When training LSTM-QR on a large set of instruments, the out-of-sample performance in a period close to the train and validation periods is superior than the benchmark models, which are recalibrated every day, as measured on the same set of instruments.

(ii) LSTM-QR can be employed on longer out-of-sample periods without the need to recalibrate.

(iii) Training LSTM-QR on a large set of instruments allows to apply it to instruments not encountered during training and achieve similar or better performance as benchmark models that were calibrated on those instruments.

We gather a list of all stocks that were constituents of the S&P 500 index at any point in time from the beginning of 1990 to the end of 2018 and download all prices in that period from the CRSP database, accessed via WRDS. We use the dividend-adjusted one-day returns calculated by CRSP, albeit converted from arithmetic to log returns. We remove all data points for which no returns were provided (which happens, e.g., on the first listing date and in case of bankruptcy) and divide the total time span into three parts. The training set is used to train the parameters. The validation set is used for the tuning of hyperparameters and to stop the training once the loss on the validation set starts increasing. The test set is then used to report statistics on the different models.[63]

To train LSTM-QR, we select a random 250 stocks that have existed during the whole period of interest. The model is then tested on those 250 stocks and another random 250 stocks. The latter 250 stocks are picked randomly from all stocks not in the training set that have prices until

[63]Due to the nature of using lagged returns the sets actually overlap for the features. As an example, consider validating using the second data point in the validation set as the target. Then, the corresponding feature set consists of the last $l-1$ data points in the training set and the first data point in the validation set. The same holds true for the test set.

the end of the testing period and at least 500 data points before the testing period starts, which is the maximum lag we allow for all models under consideration. This means that the overall test set contains the 250 stocks that LSTM-QR was trained on as well as an additional 250 stocks that it has not encountered during training.

For the numerical implementation we make use of the python library Keras[64] with the TensorFlow[65] backend. In particular, we use Keras' GPU-optimized implementation of LSTM called CuDNNLSTM. Training is performed on the Google Cloud platform, utilizing the AI Platform which also features a hyperparameter tuner based on Bayesian optimization.

3.1 Training

We train LSTM-QR just once for the case of VaR forecasting and once for iVaR forecasting. The training period spans the beginning of 1990 to the end of 2009, for validation we use data from the beginning of 2010 to the end of 2012. The benchmark models are calibrated on a daily basis.[66]

The model is trained for multiple epochs, i.e., passes through the data set. We stop training once the loss function on the validation set does not decrease for at least 100 epochs and pick the parameters of the epoch with the minimum validation loss as the final model parameters.[67] See Figure 3 for an example. This procedure prevents overfitting as the loss on the training set will generally continue to decrease as the model learns more and more specific features of that particular data set which do not generalize to unseen data.

Another common technique to reduce overfitting and thereby improving the performance on the validation and test set for RNNs is to apply a dropout layer on the inputs and/or the hidden layer.[68] A dropout layer randomly sets to zero a fraction of the units in the layer it is applied to. This can be

[64]Version 2.2.4, see Chollet et al. (2015).

[65]Version 1.12.0, see Martín Abadi et al. (2015)

[66]LSTM-QR is trained just once since training it on a similar frequency as the benchmark models is currently infeasible in the context of backtesting due to computational limitations. Comparing LSTM-QR against models that are frequently recalibrated might seem unfair. The benchmark models perform poorly when calibrated only infrequently, however, and little insight could be gained when comparing against such unrealistically constrained models. In practice, LSTM-QR can be retrained daily with less frequent hyperparameter tuning with basic cloud computing resources.

[67]The epoch with the smallest loss on the validation set is typically found after 100 to 400 epochs in our experiments.

[68]See, e.g., Srivastava et al. (2014) and Gal and Ghahramani (2016). The former suggests a dropout rate of 0.5 for the hidden units and 0.2 for the input layer for the applications they considered.

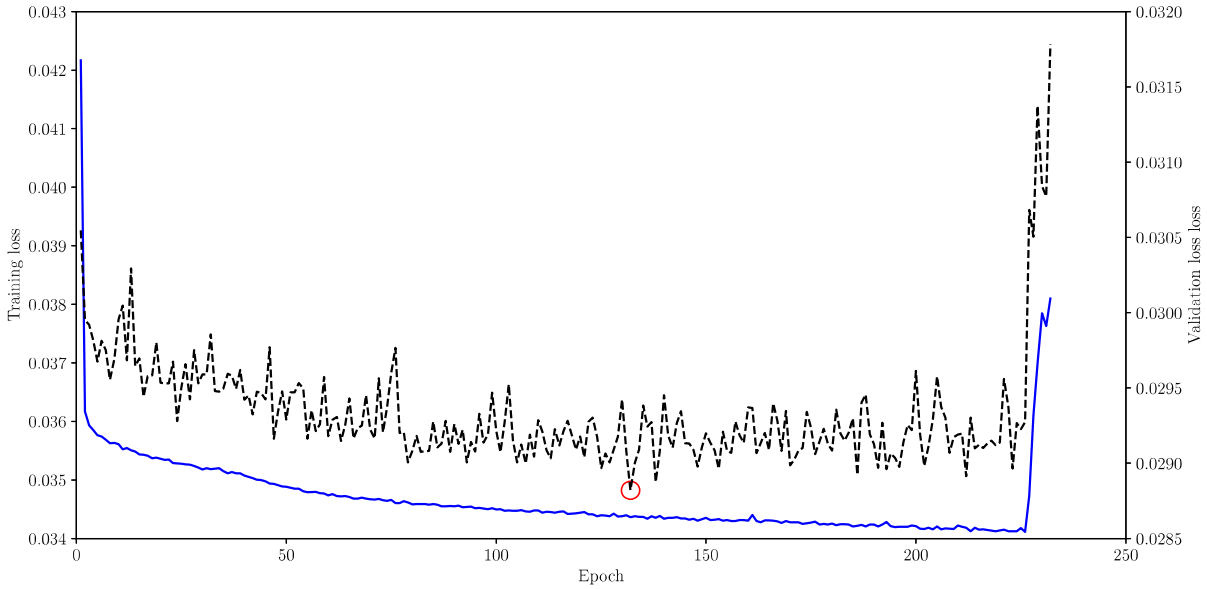


Figure 3. We plot the losses on the training set (solid blue line, primary y axis) and on the validation set (dashed black line, secondary y axis) of LSTM-QR for VaR prediction. The training loss steadily decreases except for the last few epochs where the performance deteriorates. The validation loss achieves its minimum on epoch 132, the corresponding weights of which are chosen as the model weights.

thought of as a kind of model averaging and is only applied during training. We denote the dropout rate applied to the (hidden) LSTM layer by d_h and the dropout rate applied to the input layer by d_i.

There are various hyperparameters in our model. In machine learning it is common to estimate hyperparameters by fitting a model multiple times with different values and continue with the ones that result in the lowest loss on the validation set. In many cases there are runtime restrictions that limit how many combinations of hyperparameters can be tested. Commonly applied are Bayesian methods that continuously suggest new combinations of hyperparameters that should improve the performance of the model. We rely on the hyperparameter training module offered by Google Cloud's AI Platform, which is based on Google Vizier that is used throughout Google to tune parameters.[69]

The hyperparameters that need to be estimated in our model are the number of lags, l, the number of nodes in the LSTM layer, n, the dropout rate applied to the inputs, d_i, and the dropout

[69]See Golovin et al. (2017) for more information.

rate applied to the LSTM layer, d_h.[70] Because of runtime limitations we restrict the hyperparameters to take on only discrete values $l \in \{50, 100, 200, 300, 400, 500\}$, $n \in \{10, 20, 30, 40, 50, 60, 80, 100\}$, $d_i \in \{0, 0.2\}$ and $d_h \in \{0.3, 0.4, 0.5\}$. The hyperparameters for the two cases of VaR and iVaR and the total resulting number of parameters that have to be fit are as listed in Table 1.

Case	Lag l	Nodes n	d_i	d_h	# parameters
VaR	500	50	0	0.5	11'816
iVaR	500	40	0	0.3	7'856

Table 1. Hyperparameters of LSTM-QR for the case of VaR and iVaR forecasting that minimized the respective losses on the validation set among a collection of different combinations.

While we found that 50 to 100 lags are enough to achieve acceptable results for one-step ahead forecasts, 400 to 500 are needed for the longer time horizons. In particular, for combinations of hyperparameters with $l < 400$, increasing the number of lags is the main driver for improvements and overshadows the impact of modifying all other hyperparameters. There is a much smaller difference when increasing l from 400 to 500, but still generally a slight preference for the latter. For the other hyperparameters there is a less clear trend.

Finally, we normalize the data to improve convergence. Before training starts, we collect all instrument returns available for training, remove the mean and scale to unit variance. The very same scaling is then applied to the validation and test data, which generally will not have an exact zero mean and unit variance.

3.2 Evaluation criteria

Our main goal is to develop a model that is better than the benchmarks with respect to the loss function defined in Section 2.3. The reason for this is that this loss function, coming from the theory of quantile regression, is a well-defined metric independent of the time horizon, whether returns are overlapping and the number of violations, and also provides an understanding of whether the size of the violations or non-violations are consistent with a quantile forecast. However, as most VaR models are evaluated based on a well-known set of test statistics described in Appendix B, we will also validate our model based on them. We will mention a few caveats, though.

[70] We have also tested using two or three stacked layers of LSTM units, but results on the validation set were consistently worse than using just one layer.

First, the VaR test statistics require non-overlapping returns. In order to also be able to report the VaR test statistics for $k > 1$, we choose a conservative approach. For every one of the first 1 to k points in the data set of interest, we construct non-overlapping k-period returns starting with that data point. We then evaluate the VaR test statistics on that set of data. From the collection of k test statistics calculated that way we report the worst. The reported violation rate corresponds to the set of non-overlapping data with the worst Kupiec test score. The loss, however, is calculated using overlapping returns as defined in equation (2.5).

Second, the Duration test can only be calculated if there are at least two violations. But depending on the quantile level, horizon and amount of data points, not that many violations may be expected. For that reason we exclude cases where the test cannot be computed when calculating aggregated statistics over multiple instruments.

Third and perhaps most important, the usual VaR test statistics only consider whether there was a violation or not. This can lead to degenerate cases. Consider, e.g., the case of wanting to develop a VaR model for $\alpha = 0.05$ and having 1000 returns with which to evaluate the model. Then, the VaR model

$$\mathrm{VaR}_{1-\alpha}^{\mathrm{degenerate}}(i) = \begin{cases} K & \text{if } i \,\text{modulo}\, 20 = 0 \\ -K & \text{otherwise} \end{cases}, \quad i = 1, ..., 1000 \tag{3.1}$$

would, for K large enough such that no return falls outside $(-K, K)$, result in exactly 20 evenly distributed violations. Accordingly, such a model would pass all the VaR tests of Appendix B. But for all intents and purposes, it is clearly of no use. By considering the loss function we do not judge a model based solely on whether a violation occurred, but also on the distance between the quantile forecast and the actual return.

3.3 Comparison against benchmark models

In the following we test the hypothesis that the proposed model performs better than the benchmark models on the two years after the validation set, starting from 2013-01-01 and ending at 2014-12-31, and only considering instruments with which LSTM-QR was calibrated. The aggregated statistics over all 250 instruments, all horizons and all quantile levels are displayed in Table 2a for VaR

forecasting and in Table 2b for iVaR forecasting.

For both VaR and iVaR, we find that LSTM-QR is better than all benchmarks with respect to the loss function. The VaR test statistics are comparable for all models, which can partially be attributed to the relatively short time period we are considering. We consider more detailed statistics in the following section.

Model	Loss	Kupiec	Duration	BTL
LSTM-QR	1.0157	87.80%	75.39%	79.20%
AR(1)-GARCH(1,1)-t	1.0297	89.98%	74.42%	82.08%
AR(1)-GJRGARCH(1,1,1)-t	1.0302	89.60%	74.14%	80.73%
AR(1)-GARCH(1,1,1)-st	1.0311	90.40%	74.18%	81.75%
AR(1)-GJRGARCH(1,1,1)-st	1.0324	89.58%	75.68%	79.15%
AR(1)-GARCH(1,1)-N	1.0331	86.05%	74.18%	78.62%
AR(1)-GJRGARCH(1,1,1)-N	1.0359	86.22%	75.35%	77.45%
EWMA(cal)	1.0403	87.30%	76.30%	68.65%
EWMA(0.94)	1.0569	85.82%	72.90%	62.38%
Hist	1.1248	73.42%	77.49%	90.62%

(a) VaR forecasting

Model	Loss	Kupiec	Duration	BTL
LSTM-QR	0.8911	89.12%	76.66%	73.83%
AR(1)-GJRGARCH(1,1,1)-t	0.9044	90.65%	75.00%	79.85%
AR(1)-GARCH(1,1)-st	0.9056	90.80%	75.44%	81.65%
AR(1)-GARCH(1,1)-t	0.9058	89.88%	75.68%	81.70%
AR(1)-GJRGARCH(1,1,1)-st	0.9080	90.60%	75.41%	78.75%
EWMA(cal)	0.9117	88.92%	74.43%	68.85%
AR(1)-GARCH(1,1)-N	0.9162	87.38%	76.36%	79.60%
AR(1)-GJRGARCH(1,1,1)-N	0.9175	88.25%	75.57%	77.92%
EWMA(0.94)	0.9286	87.62%	72.88%	62.72%
Hist	0.9984	71.97%	77.90%	92.17%

(b) iVaR forecasting

Table 2. Overview statistics for forecasting VaR and iVaR from 2013-01-01 to 2014-12-31 and the 250 in-sample instruments. Thereby, Loss denotes the loss averaged over the instruments and was multiplied by 10 and the Kupiec and Duration columns contain the success rates of the corresponding tests, defined as the p-values being above 5% where applicable. BTL shows the Basel Traffic Light test success rate with a test failure corresponding to a non-green zone.

3.4 Performance on out-of-sample instruments and long timescale

In the following we test the hypotheses that LSTM-QR generalizes to instruments not encountered during the training phase and to a large out-of-sample time period, without the need of recalibration.

We test all models on the period from 2013-01-01 to 2018-12-31 and both on the 250 instruments that LSTM-QR was calibrated against and 250 instruments it was not. The benchmark models are recalibrated daily on a single stock level. The aggregated statistics over all 500 instruments, all horizons and all quantile levels are displayed in Table 3a for VaR forecasting and in Table 3b for iVaR forecasting. We summarize the test statistics per horizon and quantile level in Tables 4 and 5 for VaR and iVaR, respectively, for LSTM-QR and the best benchmark model.

Model	Loss	Kupiec	Duration	BTL
LSTM-QR	1.2907	77.06%	74.60%	78.29%
AR(1)-GJRGARCH(1,1,1)-st	1.3313	78.10%	75.90%	62.55%
AR(1)-GARCH(1,1,1)-st	1.3337	77.76%	74.48%	63.05%
AR(1)-GJRGARCH(1,1,1)-t	1.3365	69.75%	75.78%	60.69%
AR(1)-GARCH(1,1)-t	1.3384	69.69%	74.81%	61.11%
EWMA(cal)	1.3583	59.29%	72.64%	51.62%
AR(1)-GARCH(1,1)-N	1.3612	58.46%	74.52%	57.75%
AR(1)-GJRGARCH(1,1,1)-N	1.3619	58.46%	76.09%	56.90%
Hist	1.3762	77.60%	67.90%	62.29%
EWMA(0.94)	1.3967	55.50%	66.81%	44.56%

(a) VaR forecasting

Model	Loss	Kupiec	Duration	BTL
LSTM-QR	1.1521	79.29%	74.76%	67.16%
AR(1)-GJRGARCH(1,1,1)-st	1.2105	75.14%	77.05%	58.65%
AR(1)-GJRGARCH(1,1,1)-t	1.2190	65.54%	77.67%	56.47%
AR(1)-GARCH(1,1)-st	1.2223	69.59%	75.98%	59.15%
AR(1)-GARCH(1,1)-t	1.2314	60.36%	76.65%	57.50%
AR(1)-GJRGARCH(1,1,1)-N	1.2369	58.23%	78.42%	53.60%
EWMA(cal)	1.2439	55.70%	73.52%	47.21%
AR(1)-GARCH(1,1)-N	1.2502	54.35%	77.92%	56.05%
EWMA(0.94)	1.2789	52.09%	65.87%	41.12%
Hist	1.2822	75.39%	66.79%	64.16%

(b) IVaR forecasting

Table 3. Overview statistics for forecasting VaR and iVaR from 2013-01-01 to 2018-12-31 and all 500 instruments. Thereby, Loss denotes the loss averaged over the instruments and was multiplied by 10 and the Kupiec and Duration columns contain the success rates of the corresponding tests, defined as the p-values being above 5% where applicable. BTL shows the Basel Traffic Light test success rate with a test failure corresponding to a non-green zone.

For both VaR and iVaR, we again find that LSTM-QR is better than all other models with respect to the loss function. We can also observe that LSTM-QR is comparable to or better than most other models when considering the Kupiec, Duration or Basel Traffic Light tests. Thereby, the differences in the Basel Traffic Light success rates are starkest, which indicates that LSTM-QR

Horizon	Quantile	Loss	VR	Kupiec	Duration	BTL
1	1%	0.6706	1.38%	74.60%	92.40%	61.40%
	5%	1.9601	4.37%	73.00%	92.40%	98.40%
	95%	1.7726	4.16%	62.80%	91.20%	99.80%
	99%	0.5810	0.86%	82.80%	91.58%	97.20%
2	1%	0.9460	0.96%	90.00%	81.01%	86.40%
	5%	2.8350	5.24%	85.60%	86.60%	83.60%
	95%	2.4965	4.30%	84.80%	82.40%	99.00%
	99%	0.8044	0.86%	82.00%	78.76%	88.80%
5	1%	1.4180	1.19%	66.00%	57.48%	68.00%
	5%	4.5409	5.39%	73.00%	70.80%	70.40%
	95%	3.9021	5.04%	81.00%	65.80%	77.60%
	99%	1.2016	0.98%	61.40%	50.19%	75.40%
10	1%	1.8871	1.14%	83.80%	51.56%	67.80%
	5%	6.3641	6.26%	72.40%	53.36%	55.60%
	95%	5.4078	5.50%	78.20%	41.94%	64.00%
	99%	1.5853	1.42%	81.60%	37.93%	59.20%

(a) LSTM-QR

Horizon	Quantile	Loss	VR	Kupiec	Duration	BTL
1	1%	0.6789	1.30%	87.80%	90.60%	71.20%
	5%	1.9832	5.48%	92.00%	97.00%	80.80%
	95%	1.7951	4.98%	96.80%	97.20%	96.40%
	99%	0.5945	1.05%	98.00%	91.80%	96.40%
2	1%	0.9839	1.67%	63.40%	84.37%	30.20%
	5%	2.8862	6.04%	79.00%	86.00%	65.20%
	95%	2.5361	5.16%	93.40%	88.20%	92.20%
	99%	0.8225	1.10%	94.40%	81.40%	82.40%
5	1%	1.5063	2.14%	43.60%	61.16%	24.40%
	5%	4.6734	7.19%	57.80%	67.80%	32.80%
	95%	3.9682	5.79%	76.80%	66.60%	63.00%
	99%	1.2266	1.13%	76.40%	47.87%	71.60%
10	1%	2.0111	2.08%	62.40%	35.20%	37.00%
	5%	6.5850	8.08%	60.00%	57.92%	31.20%
	95%	5.5065	6.64%	75.20%	47.78%	49.60%
	99%	1.6054	0.98%	92.60%	22.41%	76.40%

(b) AR(1)-GJRGARCH(1,1,1)-st

Table 4. Detailed statistics for forecasting VaR from 2013-01-01 to 2018-12-31 and all 500 instruments. Thereby, Loss denotes the loss averaged over the instruments and was multiplied by 10^3, VR the average violation rate and the Kupiec and Duration columns contain the success rates of the corresponding tests, defined as the p-values being above 5% where applicable. BTL shows the Basel Traffic Light test success rate with a test failure corresponding to a non-green zone.

Horizon	Quantile	Loss	VR	Kupiec	Duration	BTL
1	1%	0.6717	1.35%	78.00%	93.00%	61.20%
	5%	1.9610	5.36%	87.20%	89.40%	79.80%
	95%	1.7610	5.33%	91.00%	89.60%	82.00%
	99%	0.5774	1.07%	93.20%	93.20%	89.40%
2	1%	0.9162	1.53%	73.00%	86.57%	48.40%
	5%	2.6269	6.31%	55.60%	77.20%	47.00%
	95%	2.2510	4.80%	92.40%	82.60%	95.20%
	99%	0.7576	1.10%	84.40%	81.10%	78.80%
5	1%	1.3262	1.48%	69.80%	58.10%	58.60%
	5%	4.0340	6.39%	69.40%	67.40%	50.80%
	95%	3.2917	5.26%	78.40%	67.80%	71.60%
	99%	1.0931	1.48%	62.60%	58.17%	56.20%
10	1%	1.7214	1.16%	87.00%	41.79%	69.60%
	5%	5.5598	6.09%	83.00%	54.84%	61.40%
	95%	4.4263	6.14%	78.80%	50.20%	58.00%
	99%	1.3939	1.37%	84.80%	40.98%	66.60%

(a) LSTM-QR

Horizon	Quantile	Loss	VR	Kupiec	Duration	BTL
1	1%	0.6798	1.31%	85.80%	91.20%	69.20%
	5%	1.9834	5.49%	91.20%	96.20%	81.40%
	95%	1.7955	4.99%	97.20%	97.20%	96.00%
	99%	0.5949	1.06%	98.40%	92.00%	97.00%
2	1%	0.9449	1.74%	54.60%	85.17%	27.00%
	5%	2.6590	6.06%	80.00%	90.20%	65.60%
	95%	2.3050	4.95%	94.80%	87.60%	96.20%
	99%	0.7814	1.29%	91.00%	82.40%	69.80%
5	1%	1.4457	2.40%	32.40%	62.69%	16.40%
	5%	4.1516	7.25%	55.20%	72.40%	28.60%
	95%	3.3818	5.18%	85.60%	70.20%	80.40%
	99%	1.1283	1.67%	64.40%	54.43%	47.20%
10	1%	1.9454	2.79%	51.60%	44.08%	24.00%
	5%	5.8411	8.70%	53.20%	68.80%	20.00%
	95%	4.5459	5.39%	87.80%	49.70%	71.00%
	99%	1.4430	1.80%	79.00%	37.78%	48.60%

(b) AR(1)-GJRGARCH(1,1,1)-st

Table 5. Detailed statistics for forecasting iVaR from 2013-01-01 to 2018-12-31 and all 500 instruments. Thereby, Loss denotes the loss averaged over the instruments and was multiplied by 10^3, VR the average violation rate and the Kupiec and Duration columns contain the success rates of the corresponding tests, defined as the p-values being above 5% where applicable. BTL shows the Basel Traffic Light test success rate with a test failure corresponding to a non-green zone.

produces conservative forecasts as the Basel Traffic Light test can be thought of as a one-sided Kupiec test. In both cases the best benchmark model according to the loss is AR(1)-GJRGARCH(1,1,1)-st.

We compare LSTM-QR against the best benchmark in Table 4 for the case of VaR forecasting. We find that LSTM-QR has a smaller loss for every horizon and quantile level. The other test statistics do not show a clear winner between the two models. With respect to the Kupiec test LSTM-QR is worse than its benchmark for small horizons and quantile levels 95% and 99%, but better for quantile levels 1% and 5%. The Duration test is overall slightly better for the benchmark, though LSTM-QR struggles less for $k = 10$ and quantile levels 1% and 99%. However, due to the low number of expected violations the Duration test is not particularly meaningful. LSTM-QR tends to produce more conservative estimates which results in the Basel Traffic Light test succeeding more often than for its benchmark.

For the case of iVaR forecasting we compare LSTM-QR against the best benchmark in Table 5. Again, we find that LSTM-QR has a lower loss for every horizon and quantile level. The other test statistics for the one day horizon are slightly worse for LSTM-QR than for the benchmark. For longer horizons, both the Kupiec and Basel Traffic Light test tend to succeed more often for LSTM-QR than for the benchmark. For the Duration test we find the opposite, though the differences between the two models are generally smaller than for the other tests.

Figure 4 contains the histogram of differences in loss between LSTM-QR and the best benchmark model. We find that LSTM-QR is slightly better than its benchmark for most instruments, rather than have big improvements for only a few instruments. This indicates that LSTM-QR can consistently outperform its benchmark for a variety of instruments.

Finally, we show how similar VaR forecasts produced by LSTM-QR are to typical benchmark models with an example. Figure 5 displays the out-of-sample VaR forecasts of MLM (Martin Marietta Inc.) for LSTM-QR and the best benchmark. This stock was not part of the training set of LSTM-QR. We can observe that both models produce an adequate number of violations over the time frame with both Kupiec and Duration tests passing for both models. LSTM-QR features slightly less violations, closer to the ideal break rate of 5%, and the reduction in violations comes in periods of violation clusters. This can be seen around the beginning of the year 2015 where LSTM-QR produces

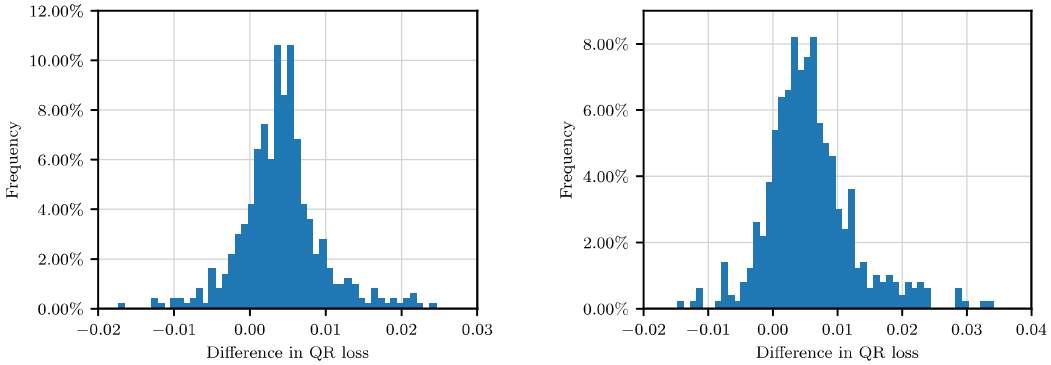


Figure 4. Histogram of differences between the loss defined in Equation (2.5) over all instruments for LSTM-QR versus the benchmark AR(1)-GJRGARCH(1,1,1)-st. A positive value means that LSTM-QR is better than the benchmark, i.e. has a smaller loss. Left: VaR forecasting. Right: iVaR forecasting.

less violations than the benchmark after the first few have occurred. LSTM-QR also seems to have less of a tendency to overreact to single large returns which can often be observed for GARCH-based models. An example of this is the large negative return at the end of the year 2015, following which the benchmark decreases the VaR forecast from -3% to -7.8%. LSTM-QR, on the other hand, decreased the forecast from -3.3% to -5.4%.

4 Extension to forecasting CVaR

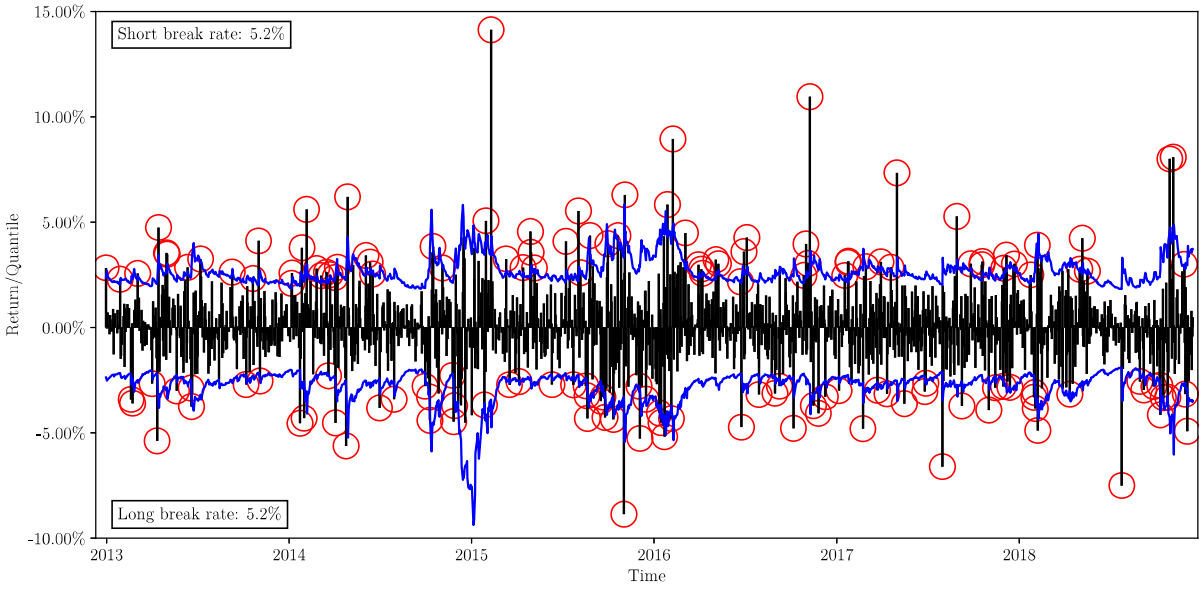
CVaR is complementing or replacing VaR as the risk measure of choice in some areas and in particular in banking due to Basel 3.5. In the following we will illustrate how the proposed model can be extended to also forecast CVaR. For that, note that a CVaR forecast for time t and level $q \leq 0.5$[71] can be represented as

$$\text{CVaR}_t^q = \frac{1}{q}\int_0^q z_t^u du. \tag{4.1}$$

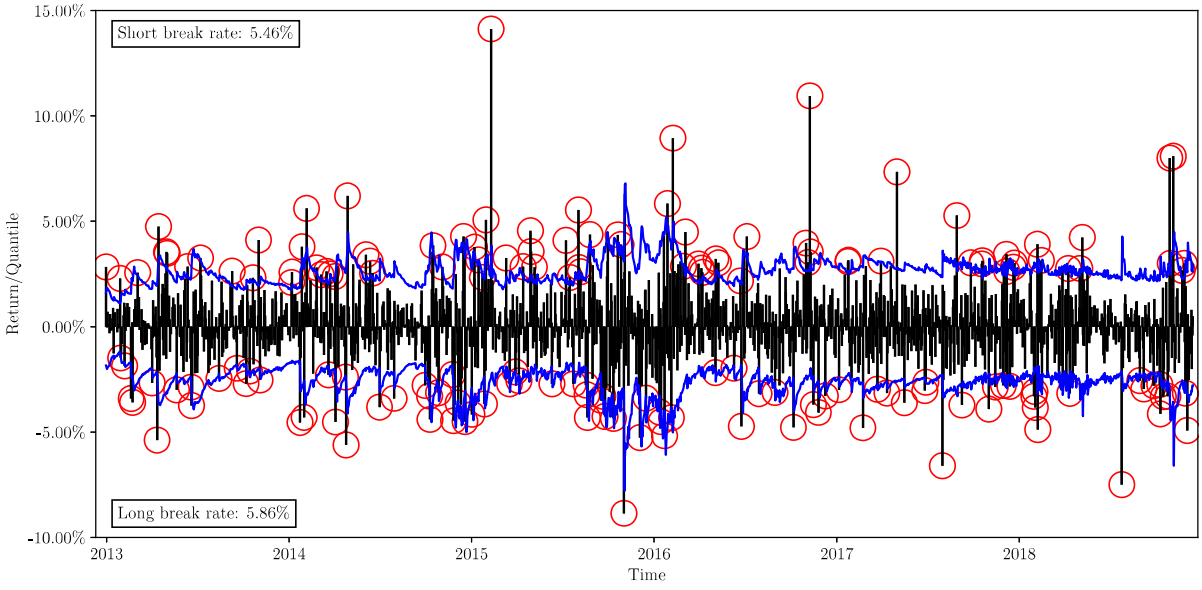
As such, assuming our model produced a continuum of quantile forecasts z_t^u for $u \leq q$ we could use equation (4.1) to create CVaR forecasts. As specified our model only produces quantile forecasts for a discrete set of quantile levels. There are multiple possible extensions that would allow the forecasting of CVaR.

First, we could modify LSTM-QR to not directly forecast quantiles, but a parametric representa-

[71] The elaborations follow similarly for $q > 0.5$.



(a) LSTM-QR



(b) AR(1)-GJRGARCH(1,1,1)-st

Figure 5. We plot the out-of-sample 5% and 95% 1-day ahead VaR forecasts from 2013-01-01 to 2018-12-31 for MLM (Martin Marietta Inc.), which was not part of the training set for LSTM-QR. Red circles indicate violations of the VaR forecast.

tion of the quantile distribution. See, e.g., Yan et al. (2018) and Gasthaus et al. (2019) for examples. The advantage of this approach is that the whole quantile (and therefore cumulative probability) distribution is available and the possibility of quantiles crossing is eliminated. The disadvantage is the possibility that the parametric representation can prove to be too restrictive.

Another approach is to make use of the loss function recently suggested by Patton et al. (2019) which is jointly minimized for VaR and CVaR. This would allow our model to simultaneously forecast VaR and CVaR for pre-determined quantile levels of interest.

Finally, we could keep the unrestricted quantile forecasts of our model and approximate the quantile distribution based on those. In the most simple case, it is even possible to approximate equation (4.1) directly via the Riemann sum

$$\text{CVaR}_t^q \approx \widehat{\text{CVaR}}_t^q = \frac{1}{q} \sum_{i=1}^{n} (u_i - u_{i-1}) z_t^{u_i} \tag{4.2}$$

for available quantile forecasts z_t^q for levels $0 < u_1 \cdots u_n = q$ and where we define $u_0 = 0$. Naturally, as $n \to \infty$, $\widehat{\text{CVaR}} \to \text{CVaR}$. This idea is exploited in, e.g., M. Kratz et al. (2018) and Couperier and Leymarie (2019) to derive implicit CVaR backtests with suggested values of n between four and six. It would also be possible to fit a parametric representation based on the quantile forecasts and then calculate CVaR via numerical integration. For the purpose of the following experiment we use the Riemann sum approximation as the same model can be used without any further modifications.

4.1 Experiment

We test CVaR forecasts for $q = 0.025$ and $q = 0.975$, which is consistent with the Basel regulation. We let LSTM-QR learn to forecast quantiles for levels $0.0025, 0.005, ..., 0.0275$ and $0.975, 0.9775, ..., 0.9975$. Hence, the CVaR approximation is based on $n = 10$ quantiles using equation (4.2).[72]

We run the same experiment as in the previous section. For the hyperparameters listed in Section 3.1 we train LSTM-QR on the same 250 instruments and the same split between training,

[72] In case of quantile crossings a common approach is to sort the quantile values and re-assign them to the quantile levels. In the Riemann approximation approach this is not necessary as we choose equally spaced levels u_i.

validation and test periods.[73] Out-of-sample performance is also again measured on the 250 training instruments and 250 other instruments. The CVaR forecasts of the best benchmark of Section 3, AR(1)-GJRGARCH(1,1,1)-st, is calculated via numerical integration based on equation (4.1) and the square-root of time rule for time horizons greater than 1.

Backtesting CVaR is a contentious and on-going topic, with some authors even suggesting it is not possible at all as CVaR is not elicitable.[74] We compare the different models based on the tests suggested in M. Kratz et al. (2018) and the statistics presented in Acerbi and Székely (2019) and Patton et al. (2019). The tests by M. Kratz et al. (2018) do not test CVaR forecasts directly, but use equation 4.2 as a justification for designing tests on the joint validity of VaR forecasts for quantile levels $u_1, ..., u_n$.[75] As our model in its basic form does not produce the whole distribution function, the test suggested in Acerbi and Székely (2019) cannot be implemented. Instead, we report the test statistic $\bar{z}_{\mathrm{ES}_\alpha}$, albeit not in terms of absolute PnL but in log-returns, which provides information about the deviation of the forecast versus realized CVaR. By fitting a parametric quantile distribution or directly forecasting one as described above, the corresponding p-values could be calculated as well. We also report the loss function L_{FZ0} as suggested in Patton et al. (2019) describing the joint fit of VaR and CVaR.

The results are listed in Table 6. As is evident, LSTM-QR is not clearly dominating the benchmark model as is the case when comparing VaR forecasts. LSTM-QR is still comparable or better than the benchmark for the two VaR-based tests Multi-Pearson and Multi-Nass for time horizons greater than 1 and especially for the 2.5% quantile level. It is slightly worse for the one-day time horizon according to these two tests. We can also observe that the loss L_{FZ0} by Patton et al. (2019) is worse for every quantile level and time horizon for LSTM-QR compared to the benchmark. The test statistic $\bar{z}$ is also mostly worse for LSTM-QR, except for time horizons > 1 and the lower quantile level.

The fact that the VaR-based tests still show acceptable or superior performance of LSTM-QR compared to the benchmark whereas the CVaR tests do not leads to the conclusion that it is not

[73] While training the hyperparameters for this specific problem could conceivably result in a better out-of-sample performance, we keep them fixed for this experiment.

[74] We refer to, e.g., Acerbi and Székely (2017) and Nolde, Ziegel, et al. (2017) for a discussion on the topic.

[75] We use $n = 5$ for the Multi-Pearson and Multi-Nass tests.

Horizon	Quantile	Loss	Pearson	Nass	$\bar{z}$
1	2.50%	-3.0733	74.20%	75.60%	0.5878%
	97.50%	-3.1942	82.60%	84.40%	0.4322%
2	2.50%	-2.7023	74.00%	76.00%	0.6193%
	97.50%	-2.8561	87.20%	88.00%	0.5842%
5	2.50%	-2.2667	79.60%	86.00%	0.5718%
	97.50%	-2.4204	69.00%	77.80%	0.7794%
10	2.50%	-1.9651	70.60%	76.40%	0.6012%
	97.50%	-2.1088	47.00%	56.00%	0.8410%

(a) LSTM-QR

Horizon	Quantile	Loss	Pearson	Nass	$\bar{z}$
1	2.50%	-3.0467	82.80%	84.20%	0.4464%
	97.50%	-3.1719	93.60%	94.40%	0.2283%
2	2.50%	-2.6601	58.40%	59.80%	0.6271%
	97.50%	-2.8341	88.80%	89.40%	0.3176%
5	2.50%	-2.1955	30.60%	43.60%	0.8013%
	97.50%	-2.4000	71.00%	79.20%	0.4106%
10	2.50%	-1.8882	27.80%	38.20%	0.8637%
	97.50%	-2.0877	56.00%	62.80%	0.5520%

(b) AR(1)-GJRGARCH(1,1,1)-st

Table 6. Detailed statistics for forecasting CVaR from 2013-01-01 to 2018-12-31 and all 500 instruments. Thereby, Loss denotes L_{FZ0} averaged over the instruments and the Pearson and Nass columns contain the success rates of the corresponding tests, defined as the p-values being above 5%. $\bar{z}$ shows the mean difference between the realized and predicted CVaR.

enough to just train a model to forecast a fixed number of quantiles without considering the overall fit of the tail of the distribution. While we did see an improvement in the CVaR tests when we increased the number of quantiles from five to ten we believe there are better and more efficient approaches to create a model for CVaR than increasing that number even further. Considering the various CVaR tests that are currently being suggested in the literature that require the knowledge of the (tails) of the distribution function, we would propose to extend the model such that it does not only forecast quantiles or CVaR, but the quantile distribution. We refer to Gasthaus et al. (2019) for an approach with a lot of potential. They suggest a spline-based approximation of the quantile distribution function, for which the number and distribution of quantile levels on which the approximation is based is learned by the model. However, as we are mainly interested in the tails of the distributions we would suggest to use a modified weighting scheme in the loss function, similar as suggested in Section 2.3. Furthermore, it may be advantageous to additionally incorporate the loss function suggested by Patton et al. (2019) to explicitly take into account the fact that the model will only forecast CVaR for certain quantile levels.

5 Conclusion

We propose a deep neural network model based on a loss function from quantile regression to forecast VaR and iVaR over multiple time horizons and quantile levels. An experiment involving 500 current and former constituents of the S&P 500 demonstrates the effectiveness of this approach, also compared to current best-practice models from the literature and industry.

The model can be applied to instruments it never encountered during training and on time periods years after the training ended. On the one hand, this can be attributed to the training scheme according to which a broad set of instruments is combined in the training set rather than training the model for every instrument. On the other hand, by producing simultaneous forecasts for a set of time horizons and quantile levels the architecture of the model allows for the training of relatively large networks as compared to the number of years available for training data. Due to both of these points, we find that just training the model once is enough to successfully apply it to six years of out-of-sample data.

While the proposed model can be applied to forecast CVaR out-of-the-box, our tests demonstrate that it may be worthwhile to extend it based on recent models suggested in the literature to forecast the whole quantile distribution and not just certain points thereof.

Future work includes modifying the loss function to take into account the clustering of violations as tested by the Duration test, forecasting the quantile distribution function and extracting information from static data such as the sector or rating of stocks and future data such as seasonality when applying the model, e.g., to the commodities market.

References

Acerbi, C., & Székely, B. (2017). General properties of backtestable statistics. *Available at SSRN 2905109*.

Acerbi, C., & Székely, B. (2019). The minimally biased backtest for ES. *Risk.net*.

Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, *9*(3), 203–228.

Bakshi, G., & Panayotov, G. (2010). First-passage probability, jump models, and intra-horizon risk. *Journal of Financial Economics*, *95*(1), 20–40.

Barone-Adesi, G., Bourgoin, F., & Giannopoulos, K. (1998). Don't look back. *Risk*, *11*, 100–103.

Basel Committee on Banking Supervision. (1996). Supervisory framework for the use of "backtesting" in conjunction with the internal models approach to market risk capital requirements. Bank for International Settlements Basel.

Bhattacharyya, M., Misra, N., & Kodase, B. (2009). MaxVaR for non-normal and heteroskedastic returns. *Quantitative Finance*, *9*(8), 925–935.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327.

Boudoukh, J., Stanton, R., Richardson, M. P., & Whitelaw, R. (2004). MaxVaR: Long-horizon value at risk in a mark-to-market environment. *Journal of Investment Management*, *2*(3), 1–6.

Brummelhuis, R., & Kaufmann, R. (2004). Time scaling for GARCH(1, 1) and AR(1)-GARCH(1, 1) processes. *Preprint, ETH Zurich*.

Chollet, F. et al. (2015). Keras.

Christoffersen, P., & Pelletier, D. (2004). Backtesting Value-at-Risk: A duration-based approach. *Journal of Financial Econometrics*, *2*(1), 84–108.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International economic review*, 841–862.

Couperier, O., & Leymarie, J. (2019). Backtesting expected shortfall via multi-quantile regression. *halshs-01909375v3*.

Engle, R. F., & Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, *22*(4), 367–381.

Föllmer, H., & Schied, A. (2011). *Stochastic finance: An introduction in discrete time.* Walter de Gruyter.

Gal, Y., & Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural networks, In *Advances in neural information processing systems.*

Gasthaus, J., Benidis, K., Wang, Y., Rangapuram, S. S., Salinas, D., Flunkert, V., & Januschowski, T. (2019). Probabilistic forecasting with spline quantile function RNNs, In *The 22nd international conference on artificial intelligence and statistics.*

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, *48*(5), 1779–1801.

Golovin, D., Solnik, B., Moitra, S., Kochanski, G., Karro, J. E., & Sculley, D. (Eds.). (2017). *Google vizier: A service for black-box optimization.*

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

J.P.Morgan. (1996). RiskMetrics - Technical Document.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33–50.

Kratz, M., Lok, Y. H., & McNeil, A. J. (2018). Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall. *Journal of Banking & Finance*, *88*, 393–407.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, *3*(2).

Leippold, M., & Vasiljevic, N. (2018). Option-implied intra-horizon value-at-risk. *Available at SSRN 2804702.*

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Jia, Y., Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, ... Xiaoqiang Zheng. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems [Software available from tensorflow.org].

Nolde, N., Ziegel, J. F. et al. (2017). Elicitability and backtesting: Perspectives for banking regulation. *The Annals of Applied Statistics*, *11*(4), 1833–1874.

Pascanu, R., Gulcehre, C., Cho, K., & Bengio, Y. (2013). How to construct deep recurrent neural networks. *arXiv preprint, arXiv:1312.6026*.

Patton, A. J., Ziegel, J. F., & Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics*, *211*(2), 388–413.

Sheppard, K. (2018). Bashtage/arch: Release 4.6.0 (version 4.6.0). https://doi.org/http://doi.org/10.5281/zenodo.1443315

Sirignano, J., & Cont, R. (2019). Universal features of price formation in financial markets: Perspectives from deep learning. *Quantitative Finance*, *19*(9), 1449–1459.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Taieb, S. B., & Atiya, A. F. (2015). A bias and variance analysis for multistep-ahead time series forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(1), 62–76.

Taylor, J. W. (2000). A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, *19*(4), 299–311.

Wen, R., Torkkola, K., & Narayanaswamy, B. (2017). A multi-horizon quantile recurrent forecaster. *arXiv preprint, arXiv:1711.11053*.

Xu, Q., Liu, X., Jiang, C., & Yu, K. (2016). Quantile autoregression neural network model with applications to evaluating value at risk. *Applied Soft Computing*, *49*, 1–12.

Yan, X., Zhang, W., Ma, L., Liu, W., & Wu, Q. (2018). Parsimonious quantile regression of financial asset tail dynamics via sequential learning, In *Advances in neural information processing systems*.

Zumbach, G. (2007). The RiskMetrics 2006 methodology.

Appendices

A Benchmark models

A.1 Historical model

The first benchmark model was also the first to be used historically to measure VaR. For VaR estimates, we estimate the one-step ahead quantile as the sample quantile,

$$z_{t+1}^q = \text{sample quantile}_q(x_{t-l+1}, ..., x_t). \tag{A.1}$$

Multi-step quantiles could be estimated by considering non-overlapping multi-step returns. However, for 10 day returns and 500 data points this results in a sample of only 50 returns to estimate a quantile of up to 99%. We therefore resort to the square-root of time rule, i.e., $z_{t+k}^q = \sqrt{k} z_{t+1}^q$.

IVaR estimates z_{t+k}^q are created by sampling $N = 100'000$ sets of k daily returns from $x_{t-l+1}, ..., x_t$. We then calculate the minimum (for $q \leq 0$), respectively, maximum (for $q > 0$) cumulative return from 1 to k to create a distribution of N minimum, respectively, maximum k-period returns. Then we take the sample quantile of that distribution.

A.2 Variance and mean models

The second set of benchmarks all model the variance and the mean process and create quantile forecasts based thereupon. In particular, assume returns are modeled as

$$x_t = \mu_t + \varepsilon_t, \tag{A.2}$$

for some mean process $\mu = (\mu_t)_t$, where the residuals are assumed to follow

$$\varepsilon_t = \sigma_t z_t \tag{A.3}$$

for some variance process $\sigma^2 = (\sigma_t^2)_t$ and z_t some standard strong white noise process.

For VaR estimates, the quantiles are estimated using the inverse distribution of the process z_t,

denoted by F^{-1}, and the one-step forecasts of both μ_t and σ_t, i.e.

$$z_{t+1}^q = \mu_{t+1} + \sigma_{t+1} F^{-1}(q). \tag{A.4}$$

For time horizons greater than one we apply the square-root rule, i.e., $z_{t+k}^q = \sqrt{k} z_{t+1}^q$.[76]

For iVaR estimates, we simulate N paths of x from t to $t+k$ and create the distribution of the minimum (for $q \leq 0$), respectively, maximum (for $q > 0$) cumulative returns of which we take the sample quantile.[77] For the numerical implementation of these models we make use of the python package ARCH.[78]

(GJR-) GARCH models

The first sub-class of benchmark models consists of the classical GARCH[79] and GJR-GARCH[80] models, combined with an auto-regressive model for the mean and different distributional assumptions for the residuals.

The mean is given by an AR(l) model

$$\mu_t = c + \sum_{i=1}^{l} \varphi_i x_{t-i} \tag{A.5}$$

with parameters c and φ_i. The variance σ_t^2 is modeled as a GJR-GARCH(p, o, r) process[81]

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{o} \gamma_j \varepsilon_{t-j}^2 \mathbb{1}_{\{\varepsilon_{t-j}<0\}} + \sum_{k=1}^{r} \beta_k \sigma_{t-k}^2 \tag{A.6}$$

with parameters ω, α_i, β_i, γ_i. The innovations z_t are assumed to be distributed with the normal, Student's t or the skewed Student's t distribution. In all cases the distributions are normalized to produce zero mean and unit variance random variables. The models are calibrated using standard

[76] Alternatively, we could simulate paths as for the iVaR estimation. The significantly higher computational cost and associated sampling error outweigh the disadvantages of using the square-root rule approximation.

[77] We chose $N = 10'000$ for providing a reasonable trade-off between computational time and simulation error.

[78] See Sheppard (2018).

[79] See Bollerslev (1986)

[80] See Glosten et al. (1993)

[81] GARCH is a special case of GJR-GARCH via GARCH(p, r) = GJR-GARCH($p, 0, r$).

maximum likelihood techniques.

We restrict our attention to the cases with only one lag. We have also considered using zero or constant mean assumptions, but did not find them to influence the results significantly.

The parametrization of these benchmark models are denoted as AR(1)-GJRGARCH(1,1,1)-N/t/st, and AR(1)-GARCH(1,1,1)-N/t/st, where N, t and st refer to the normal, Student's t and skewed Student's t-distributions, respectively.

EWMA

The EWMA model, also called the RiskMetrics™ 1996 model, as proposed by J.P.Morgan (1996) assumes that the mean is zero, $\mu \equiv 0$, that the variance process is given by,

$$\sigma_t^2 = \lambda \sigma_{t-1}^2 + (1-\lambda)\varepsilon_{t-1} \tag{A.7}$$

and that z_t is distributed standard-normally.[82]

While it was originally proposed to set the parameter λ to a fixed value, often 0.94 for one day forecasts, we also consider the case of calibrating λ using maximum likelihood.

The parametrization of these benchmark models are denoted as EWMA(λ) and EWMA(cal) for the case for a fixed $\lambda > 0$ or a daily calibrated λ, respectively.

B Value-at-Risk tests

For VaR or iVaR estimate z_{t+k}^q, for some time horizon k and quantile level q, we say that out-of-sample return x_{t+k} is a violation if it is smaller than (for $q \leq 0.5$) or greater than (for $q > 0.5$) z_{t+k}^q.

B.1 Kupiec test

Historically the first test that was broadly used to judge the goodness of VaR models, the test suggested by Kupiec (1995) judges whether the number of violations are consistent with the quantile

[82]The variance process is a special case of an IGARCH model, often also called EWMA.

level q. The null hypothesis is that the number of violations equals the quantile level q.

Under the assumption that forecasts are independent, the probability of observing n violations over N returns is given by the binomial distribution. This suggests a log likelihood ratio test with the test statistic

$$\Lambda = -2\log\left(\frac{q^n(1-q)^{N-n}}{\left(\frac{n}{N}\right)^n\left(1-\frac{n}{N}\right)^{N-n}}\right). \tag{B.1}$$

Under the null hypothesis Λ is chi-square distributed, $\Lambda \sim \chi(1)^2$.

The null hypothesis is typically rejected, and the test therefore failed, if the p-value of the test statistic is smaller than 0.05.

B.2 Duration test

Besides violations occurring in a number that is consistent with the quantile level, as validated by the Kupiec test, the time between violations should also be independent of the past and not cluster. Otherwise, a VaR model may for example pass the Kupiec test for a certain time period because there were too many violations at the beginning of the period that were canceled out by there not being enough in the rest of the period. The first test for the timing of violations was suggested by P. F. Christoffersen (1998), which explicitly assumes violations are independent and was found to have little power in real-world applications. In the following we present the Duration test by P. Christoffersen and Pelletier (2004).

The duration between violations should be i.i.d. To construct a test for this, consider that the only continuous-valued distribution that is memoryless is the exponential distribution with pdf $f_e(x,\lambda) = \lambda e^{-\lambda x}$, $\lambda > 0$. That distribution is a special case of the Weibull distribution, with pdf

$$f_W(x,a,b) = a^b b x^{b-1} - (ax)^b, \quad a,b > 0 \tag{B.2}$$

for $b = 1$. We can therefore fit a Weibull distribution to the durations of violations and have the null hypothesis that $b = 1$. Note that under the null hypothesis, the average time between violations is $1/q$ days for quantile level q.

We denote by $t_1, ..., t_{\widetilde{N}}$ the times of violations in the period $\{1, ..., T\}$. The duration of these

violations is given by $\widetilde{D}_i = t_{i+1} - t_i$ for $i = 1, ..., \widetilde{N} - 1$. Since we want to include all the information offered by the timing of the first and last violation, special care needs to be taken if they did not happen at the beginning and end of the time period as they will then be censored observations. As such we define the duration including censored data, $D_1, ..., D_N$, for $i = 1, ..., \widetilde{N} - 1$ as

$$\begin{aligned} D_i &= \widetilde{D}_i & t_1 &= 1, t_N = T \\ D_1 &= t_1, D_{i+1} = \widetilde{D}_i & t_1 &> 1, t_N = T \\ D_i &= \widetilde{D}_i, D_N = T - t_{\widetilde{N}} & t_1 &= 1, t_N < T \\ D_1 &= t_1, D_{i+1} = \widetilde{D}_i, D_N = T - t_{\widetilde{N}} & t_1 &> 1, t_N < T \end{aligned} \tag{B.3}$$

We also define the censor flags

$$C_1 = \begin{cases} 0 & , t_1 = 1 \\ 1 & , \text{ otherwise} \end{cases}, \quad C_N = \begin{cases} 0 & , t_{\widetilde{N}} = T \\ 1 & , \text{ otherwise} \end{cases}. \tag{B.4}$$

The likelihood consists of three parts. The contribution of the first observation, of the last observation and of the rest, which are by construction uncensored. If D_i is censored, then its contribution is not given by its probability density function f but its survival function S. As such we have

$$\begin{aligned} \log L(D, \theta) = C_1 \log S(D_1, \theta) + (1 - C_1) \log f(D_1, \theta) + \sum_{i=2}^{N-1} \log(f(D_1, \theta)) \\ + C_N \log S(D_N, \theta) + (1 - C_{D_N}) \log f(D_N, \theta) \end{aligned}$$

for parameters θ.

The log likelihood ratio test statistic Λ is then given by

$$\Lambda = -2(L_e(D, \theta_e) - L_W(D, \theta_W)) \tag{B.5}$$

where θ_e and θ_W give the maximum likelihood for the exponential and Weibull distribution, respectively. Λ is asymptotically $\chi(1)^2$ distributed under the null distribution.

The null hypothesis is typically rejected, and the test therefore failed, if the p-value of the test

statistic is smaller than 0.05.

B.3 Basel Traffic Light test

The Basel Traffic Light test, as introduced by Basel Committee on Banking Supervision (1996), can be thought of as a one-sided version of the Kupiec test, with only too many violations leading to a test failure. For n violations out of a sample of N data points with quantile level q there are three categories,

$$C(n,N,p) = \begin{cases} \text{green} & , \text{ if } B(n,N,p) < 0.95 \\ \text{yellow} & , \text{ if } 0.95 < B(n,N,p) < 0.9999 \\ \text{red} & , \text{ if } B(n,N,p) > 0.9999 \end{cases}, \tag{B.6}$$

where $B(n,N,p)$ is the cumulative binomial distribution function. In our benchmark analysis, we denote an instrument as having failed the Basel Traffic Light test if the category $C(n,N,p)$ is yellow or red.

C Loss function - quantile level multiplier

For quantile levels $q \in \{0.01, 0.05, 0.95, 0.99\}$, we defined the multiplier

$$m^{\text{level}}(q) = \begin{cases} 3 & \text{if } q \in \{0.01, 0.99\} \\ 1 & \text{otherwise} \end{cases} \tag{C.1}$$

to ensure that $m^{\text{level}}(q) L_q(x^q_{t+k}, z^q_{t+k})$ are of similar size for all q, assuming the same time horizon k.

To justify this from a theoretical point of view, assume x is a random variable of which we know the true quantile z_q at level q. Then, the expected value of the loss function is

$$\mathbb{E}[L_q(x,z_q)] = (1-q)q(\mathbb{E}[x \,|\, x \geq z_q] - \mathbb{E}[x \,|\, x \leq z_q]) = (1-q)q(\text{CVaR}_q(-x) - \text{CVaR}_{1-q}(x)). \tag{C.2}$$

As such, we can calculate that $\frac{\mathbb{E}[L_q(x,z_{0.05})]}{\mathbb{E}[L_q(x,z_{0.01})]}$ is 3.87 when assuming x follows a normal distribution and for a Student's t-distribution the ratio ranges from 2.77 (three degrees of freedom) to 3.87

(limiting case, degrees of freedom going to infinity), independent of the mean and volatility of the respective distributions. Due to the symmetry of these distributions, these ratios also hold for the other combinations of the quantile levels (0.05 to 0.99, 0.95 to 0.01, 0.95 to 0.99). We can also observe empirically that these ratios are also usually between 2.5 and 3.5 for the benchmark models considered in this paper on various sets of real or artificial data, time periods and time horizons.

Therefore, we propose to use the multiplier $m^{\text{level}}(q)$ as defined above. However, should it turn out during the training phase that certain quantile levels are consistently fit better than others, this multiplier as well as the horizon multiplier should be adjusted to ensure a comparably good fit over all quantile levels and horizons.

Part III

Appendix

Curriculum Vitae

Personal Data

Full Name	Steven Patrick Schärer
Date of Birth	14 January 1990

Education

Sep 2014 - Jul 2020	PhD Student at Department of Banking & Finance, University of Zurich, Switzerland
Sep 2012 - Apr 2014	Master of Science in Quantitative Finance, University of Zurich & ETH Zurich, Switzerland
Sep 2008 - Mar 2012	Bachelor of Science in Mathematics, ETH Zurich, Switzerland

Work Experience

Jun 2019 - present	Quant Strats, Credit Suisse AG, Zurich, Switzerland
Mar 2014 - May 2019	Senior Quant Engineer, swissQuant Group AG, Zurich, Switzerland
May 2012 - Feb 2014	Performance Measurement Analyst, UBS AG, Zurich, Switzerland